CONTRIBUTIONS TO PARAMETRIC STATISTICAL THEORY AND PRACTICE

JOHN AITCHISON

Volume 1

D.Sc.
University of Edinburgh
1980

To M, the constant among many variables

It is sometimes considered a paradox that the answer depends not only on the observations but on the question: it should be a platitude.

H. Jeffreys

ACKNOWLEDGEMENTS

I wish to express my grateful thanks to a large number of friends.

To all the non-statisticians who over the last thirty years have generously shared their encounters with variability and uncertainty and so acted as a stimulus to my statistical research.

To my colleagues in statistical research in the Universities of Cambridge, Glasgow, Liverpool and Hong Kong, who in daily contact have helped to clarify ideas more than they realise.

To the many students, in undergraduate study and research, who by their unwillingness to accept dogma have often rocked the statistical boat.

To all the secretaries who with skill and good humour have turned unreadable manuscript into beautiful typescript; and in particular to Mrs Lucile Y.K. Lo who, in addition to typing the commentary and many of the papers, has helped in the collation of the material presented.

J. Aitchison.

ABSTRACT

Published work (twenty research papers and two books) and two unpublished papers are presented within the context of a short commentary on the theme of statistical parametric modelling. A secondary theme is the stimulus brought to statistical theory through close attention to the requirements of particular practical problems.

The developments discussed cover the main divisions of parametric modelling: model selection, model validity, estimation, hypothesis testing, experimental design, prediction, decision making, model fitting and complex modelling. In model selection and validity the presentation begins and ends with an intensive study of two important particular classes of models, the centenarian lognormal class and a new logistic-normal class with wide application in the analysis of compositional and probabilistic data. In estimation and hypothesis testing the main aim is the provision of routine methodology, to allow the easy consideration of non-standard and complex situations. The discussion includes multiple hypothesis testing problems, in particular the use of restricted and confidence-region tests, and a problem of constructing optimum designs for certain comparative trials. The emergence of statistical prediction analysis and the central role of predictive distributions as an important tool in many practical situations, with advantages of realism and tractability over other methods, are explained. The necessary theory for applications in medicine, in particular to statistical diagnosis, is developed, and the more complex models required to take account of difficulties of calibration, imprecision and uncertain diagnoses are constructed. Methods of comparing human inferential judgment against statistical modelling are demonstrated, and the possibility of estimating the implicit utility functions used by clinicians in their allocations of treatments to patients is explored.

CONTENTS

Volume 1

Preface	*page*	1
List of published works submitted		3
List of unpublished works submitted		6
Commentary (including details of sections and papers submitted)		7
References		433

Volume 2

Two published books

(a) The Lognormal Distribution	*page*	1
(b) Statistical Prediction Analysis		192

PREFACE

The works submitted in support of this candidature for the degree of D.Sc. are set out in chronological order in the lists immediately following this Preface. To help the reader to relate these twenty-four works to the theme of this submission I provide a short commentary tracing the practical motivation for the statistical concepts and principles introduced, outlining the development of appropriate statistical methodology, and indicating the relevance to other work. Again for the convenience of the reader, the twenty published and the two unpublished *papers* are reproduced in a standard xerographic format at appropriate places within the commentary rather than presented as an untidy appendix of differently sized reprints: these form volume 1. The two *books* presented have also been reproduced in xerographic format and form volume 2.

In the commentary I have found it convenient to use the (Authors, Year) form of reference but in order to distinguish between submitted and other published works I have also quoted for any submitted work its serial number in the list of submitted works. Thus Aitchison (6:1965) refers to submitted paper no. 6, whereas Aitchison and Silvey (1957) specifies a published though not submitted paper, details of which are contained in the list of references at the end of the first volume.

The main theme of this presentation is parametric statistical theory and practice, in particular the formulation of practical problems in terms of parametric models and the development of new statistical concepts, principles and methodology towards a sensible

resolution of these problems. A secondary, but in my view equally important, theme is the modest reiteration of a tenet, well expressed by Chebyshev in the Drawing of Maps, that 'the bringing together of theory and practice leads to the most favourable results: not only does practice benefit, but the sciences themselves develop under the influence of practice, which reveals new subjects for investigation and new aspects of familiar subjects.'

This secondary theme stems from a realisation that all the statistical research presented here has arisen out of the needs of specific consultative problems for a variety of departments in the Universities of Cambridge, Glasgow, Hong Kong and Liverpool and for teaching hospitals in Glasgow, Guangchou and Hong Kong. It is perhaps surprising that attention to the demands of the variety of practical problems which arrive on a statistician's desk in a more or less arbitrary sequence should lead to the development of a statistically coherent body of knowledge rather than a hotchpotch of unrelated statistical exercises. The emergence of satisfactory mathematical statistical theory does, however, seem to quicken when, according to Chebyshev's dictum, the statistician travels hand in hand with practice. While the extent to which this has been satisfactorily achieved for the particular problems posed by my consultees can only judged by them, I hope that the statistical content of the works submitted, however theoretical some may appear in isolation, may be seen by the reader through the help of the commentary to be sensible means towards very practical ends.

John Aitchison

LIST OF PUBLISHED WORKS SUBMITTED

1 AITCHISON, J. and BROWN, J.A.C. (1957)

The Lognormal Distribution

Cambridge University Press (176 pages)

2 AITCHISON, J. and SILVEY, S.D. (1958)

Maximum-likelihood estimation of parameters subject to restraints

Ann. Math. Statist. 29, 813-28

3 AITCHISON, J. and SILVEY, S.D. (1960)

Maximum-likelihood estimation procedures and associated tests of significance

J.R. Statist. Soc. B22, 154-71

4 AITCHISON, J. (1961)

The construction of optimal designs for the one-way classification analysis of variance

J.R. Statist. Soc. B23, 352-67

5 AITCHISON, J. (1962)

Large-sample restricted parametric tests

J.R. Statist. Soc. B24, 234-50

6 AITCHISON, J. (1964)

Confidence-region tests

J.R. Statist. Soc. B26, 462-76

7 AITCHISON, J. (1964)

Bayesian tolerance regions

J.R. Statist. Soc. B26, 161-75

8 AITCHISON, J. (1965)

Likelihood-ratio and confidence-region tests

J.R. Statist. Soc. B27, 245-50

9 AITCHISON, J. and SCULTHORPE, Diane (1965)

Some problems of statistical prediction

Biometrika 52, 469-83

10 AITCHISON, J. (1966)

Expected-cover and linear-utility tolerance intervals

J.R. Statist. Soc. B28, 57-62

11 AITCHISON, J. (1970)

Statistical problems of treatment allocation

J.R. Statist. Soc. A133, 206-28

12 AITCHISON, J. and BENNETT, J.A. (1970)

Polychotomous quantal response by maximum indicant

Biometrika 57, 253-62

13 AITCHISON, J. and DUNSMORE, I.R. (1975)

Statistical Prediction Analysis

Cambridge University Press (273 pages)

14 AITCHISON, J. (1975)

Goodness of prediction fit

Biometrika 62, 547-54

15 AITCHISON, J. and KAY, J.W. (1975)

Principles, practice and performance in decision-making in clinical medicine

In *The Role and Effectiveness of Decision Theories in Practice* (eds K.C. Bowen and D.J. White): London: Hodder and Stoughton, pp.252-72

16 AITCHISON, J. and BEGG, C.B. (1976)

Statistical diagnosis when basic cases are not classified with certainty

Biometrika 63, 1-12

17 AITCHISON, J., HABBEMA, J.D.F. and KAY, J.W. (1977)

A critical comparison of two methods of statistical discrimination

Applied Statistics 26, 15-25

18 AITCHISON, J. (1977)

A calibration problem in statistical diagnosis: the system transfer problem

Biometrika 64, 461-72

19 AITCHISON, J. (1979)

A calibration problem in statistical diagnosis: the clinic amalgamation problem

Biometrika 66, 357-66

20 AITCHISON, J. and LAUDER, I.J. (1979)

Statistical diagnosis from imprecise data

Biometrika 66, 475-83

21 AITCHISON, J. (1979)

Calibration and assay from imprecise data.

Bull. Inst. Int. Statist. 48, 4, 9-12

22 AITCHISON, J. and SHEN, S.M. (1980)

Logistic-normal distributions: some properties and uses

Biometrika 67

LIST OF UNPUBLISHED WORKS SUBMITTED

23 AITCHISON, J. (1980a)

A new approach to null correlations of proportions

Submitted to *J. Math. Geol.*

24 AITCHISON, J. (1980b)

Testing for additive isometry and proportional invariance

Submitted to *Biometrics*

COMMENTARY

1 INTRODUCTION: PARAMETRIC STATISTICAL MODELLING *page* 11

2 LOGNORMAL MODELS 13

Aitchison, J. and Brown, J.A.C. (1957) *in vol.* 2

The Lognormal Distribution

Cambridge University Press

3 GENERAL TECHNIQUES OF PARAMETRIC ESTIMATION AND HYPOTHESIS TESTING *page* 20

Aitchison, J. and Silvey, S.D. (1958) 27

Maximum-likelihood estimation of parameters subject to restraints

Ann. Math. Statist. 29, 813-28

Aitchison, J. and Silvey, S.D. (1960) 44

Maximum-likelihood estimation procedures and associated tests of significance

J.R. Statist. Soc. B22, 154-71

4 MULTIPLE HYPOTHESIS TESTING 63

Aitchison, J. (1962) 71

Large-sample restricted parametric tests

J.R. Statist. Soc. B24, 234-50

Aitchison, J. (1964) 89

Confidence-region tests

J.R. Statist. Soc. B26, 462-76

Aitchison, J. (1965) 105

Likelihood-ratio and confidence-region tests

J.R. Statist. Soc. B27, 245-50

5 CONSTRUCTION OF F-OPTIMAL DESIGNS *page* 112

Aitchison, J. (1961) 114

The construction of optimal designs for the one-way classification analysis of variance

J.R. Statist. Soc. B23, 352-67

6 PARAMETRIC TOLERANCE REGIONS 131

Aitchison, J. (1964) 135

Bayesian tolerance regions (with discussion)

J.R. Statist. Soc. B26, 161-72 and 192-210

7 PREDICTIVE DENSITY FUNCTIONS 170

Aitchison, J. and Sculthorpe, Diane (1965) 172

Some problems of statistical prediction

Biometrika 52, 469-83

Aitchison, J. (1966) 188

Expected-cover and linear-utility tolerance intervals

J.R. Statist. Soc. B28, 57-62

8 SOME PROBLEMS OF DECISION MAKING 195

Aitchison, J. (1970) 198

Statistical problems of treatment allocation (with discussion)

J.R. Statist. Soc. A133, 206-28 and 229-38

Aitchison, J. and Bennett, J.A. (1970) 232

Polychotomous quantal response by maximum indicant

Biometrika 57, 253-62

9 STATISTICAL PREDICTION ANALYSIS 243

Aitchison, J. and Dunsmore, I.R. (1975) *in vol.* 2

Statistical Prediction Analysis

Cambridge University Press

10 ESTIMATIVE AND PREDICTIVE MODEL FITTING *page* 248

Aitchison, J. (1975) 253

Goodness of prediction fit

Biometrika 62, 547-54

Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977) 262

A critical comparison of two methods of statistical discrimination

Applied Statistics 26, 15-25

11 THE ANALYSIS OF SUBJECTIVE PERFORMANCE IN INFERENTIAL TASKS 274

Aitchison, J. and Kay, J.W. (1975) 277

Principles, practice and performance in decision-making in clinical medicine

In *The Role and Effectiveness of Decision Theories in Practice* (eds K.C. Bowen and D.J. White). London: Hodder and Stoughton, pp.252-72

12 PARAMETRIC MODELLING IN STATISTICAL DIAGNOSIS 299

Aitchison, J. and Begg, C.B. (1976) 309

Statistical diagnosis when basic cases are not classified with certainty

Biometrika 63, 1-12

Aitchison, J. (1977) 322

A calibration problem in statistical diagnosis: the system transfer problem

Biometrika 64, 461-72

Aitchison, J. (1979) 335

A calibration problem in statistical diagnosis: the clinic amalgamation problem

Biometrika 66, 357-66

Aitchison, J. and Lauder, I.J. (1979) *page* 346

Statistical diagnosis from imprecise data

Biometrika 66, 475-83

Aitchison, J. (1979) 356

Calibration and assay from imprecise data

Bull. Inst. Int. Statist. 48, 4, 9-12

13 LOGISTIC-NORMAL MODELS 361

Aitchison, J. and Shen, S.M. (1980) 369

Logistic-normal distributions: some properties and uses

Biometrika 67

Aitchison, J. (1980a) 382

A new approach to null correlations of proportions

Submitted to *J. Math. Geol.*

Aitchison, J. (1980b) 411

Testing for additive isometry and proportional invariance

Submitted to *Biometrics*

14 CONCLUSION 431

1 INTRODUCTION: PARAMETRIC STATISTICAL MODELLING

Suppose that for some experiment or observational situation f under investigation we have adopted a statistical description with record set Y and a class of possible probability models, say density functions $p(y|\theta)$ on Y, where θ is an indexing parameter belonging to a finite-dimensional parameter set Θ. The true indexing parameter θ^* is unknown. Suppose further that we have obtained, or could obtain, data x from another related experiment or observational situation e with record set X, with describing class of density functions $p(x|\theta)$ indexed by parameters in the same parameter set Θ and with the same true, but unknown, parameter θ^*. For example, a common situation is where e is n replicates of f, providing data $x = (x_1, \ldots, x_n)$ on which to base inferences about f.

From such situations there arises a whole range of statistical problems of which the following are some of the most important.

1. *Model selection.* The problem of selecting an appropriate parametric form for $p(y|\theta)$ and consequently for $p(x|\theta)$.
2. *Model validity.* Testing the validity of the form selected from the information in x.
3. *Estimation.* Using x to obtain information on, or to estimate, θ^*.
4. *Hypothesis testing.* Using x to test some prestated hypothesis concerning θ^*.
5. *Experimental design.* When there is a class E of possible 'informative' experiments available, the problem of how to choose e from E to serve best the practical purpose of the investigation.

6. *Prediction.* Using x to make some statement about, or to predict, the outcome y of the experiment f.

7. *Decision making.* Relating the modelling to some decision-making problem and using x to determine an appropriate course of action.

8. *Model fitting.* Using x to assess the complete density function of f.

9. *Complex modelling.* Using simple parametric models as components in the building of a model of a more complex system.

We shall see how all these aspects arise, some repeatedly, in the course of this commentary on the works submitted. Our first concern is an intensive study of one important class of parametric models.

2 LOGNORMAL MODELS

2.1 A monograph on a single distribution

In most disciplines where variability and uncertainty play a significant role workers almost inevitably encounter some data whose pattern of variability is adequately described by a lognormal model, with density function of the form

$$p(y|\mu,\sigma) = \frac{1}{y\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{(\log y - \mu)^2}{2\sigma^2}\right\} \qquad (y > 0),$$

or its higher-dimensional counterpart. Such models have been known and applied repeatedly since their first explicit recognition and development a century ago by McAlister (1879). In the period 1952-56 when Aitchison and Brown (1954a,b,c) were collaborating in studies of demand analysis, income distribution and the analysis of family budgets they realised that some of the necessary lognormal distribution results which they had independently developed had already been discovered, and indeed often rediscovered many times. Their own experience of such rediscovery, and the realisation that others in the future might unnecessarily tread the same well-worn path, had suggested the undertaking of a collation of the diffuse literature on the theory and application of lognormal models. Almost predictably the project grew beyond review paper dimension to monograph size, as the extent of the previous literature was realised, as deficiences in existing theory emerged, as under-developed or new properties had to be investigated and as research into economic applications progressed.

The resulting monograph, Aitchison and Brown (1:1957) was at

the date of its publication largely experimental, the first substantial work devoted to a single form of distribution. Some statisticians expressed doubts as to the wisdom of such a venture. That the monograph has been followed by others such as Haight (1967) on Poisson models, Lancaster (1969) on the chi-squared distribution and Ashton (1972) on logistic models, together with a procession of encyclopaedic volumes having substantial chapters on individual distributions such as the revived Elderton and Johnson (1969), Johnson and Kotz (1969, 1970, 1972), Mardia (1970), Ord (1972), Patil, Kotz and Ord (1975), is some evidence in support of the experiment.

Claims to original or significant contribution to knowledge cannot, of course, be based on collative ability and constant, though modest, sales and so what may be regarded as such contributions are now indicated.

1. The generalisation of the moment distribution property (§2.8).
2. A clear statement of, and some variations on, models of genesis (Chapter 3).
3. An extensive study of the relative merits of different estimation methods for the two- and three-parameter models, including, at that time unusual, assessment by simulation (Chapters 4,5,6).
4. A quantitative response model with multiplicative lognormal error, allowing for heteroscedasticity (§7.9).
5. The relevance of lognormal models to the statistical analysis of income distributions, in particular the role of moment distributions in Lorenz curve analysis and of the σ parameter in measures of concentration (Chapter 11).

6. The use of lognormal models in the statistical analysis of consumer behaviour, in particular the synthesis of Engel curve theory and the role of the lognormal convolution property in aggregation problems of demand analysis (Chapter 12).

2.2 More recent development

In the twenty-three years since the publication of the monograph there appear to have been few new theoretical results of any apparent significance.

1. Heyde (1963) shows that a lognormal distribution is not uniquely determined by its moments, a result of considerable theoretical interest, though not of any practical significance.
2. Hill (1963) demonstrates that the likelihood function associated with a sample from a three-parameter lognormal model tends to ∞ as the parameter vector (τ, μ, σ), in the notation of Aitchison and Brown (1:1957), tends along a certain path to $(x_0, -\infty, \infty)$, where x_0 is the minimum sample value. Thus the clearly ridiculous estimate $(x_0, -\infty, \infty)$ of (τ, μ, σ) is a maximum likelihood estimate, providing the absolute maximum ∞ of the likelihood. This unfortunate feature of maximum likelihood estimation had gone unnoticed by earlier writers, including Aitchison and Brown (1:1957), who mistakenly identified relative or local maximum likelihood estimators as providing the absolute maximum of the likelihood function. This feature is theoretically irritating rather than practically awkward since the relevance of the local maximum to most practical problems is readily restored by the mildest of prior assumptions about the parameters, giving negligible prior probability to parameter vectors in this absurd region at infinity.

3. Since 1957 there have been two streams of development in testing model validity which provide extra opportunities for testing the lognormality of data. The first stream has produced new tests of univariate and multivariate normality, and hence of lognormality: see for example, Box and Cox (1964), Shapiro and Wilk (1965), Healy (1968), Andrews, Gnanadesikan and Warner (1973), Cox and Small (1978). The second stream has been concerned with problems of testing separate classes of models, and so allows the testing of a lognormal model against specific parametric alternatives such as the exponential model and the gamma model: see, for example, Cox (1961, 1962), Atkinson (1969, 70).

4. Mosimann (1970, 1975a,b) gives some characterisations of multivariate lognormal distributions in relation to biometric analysis of shape and size. One characterisation in particular has played an important deterrent role against the use of lognormal distributions in such analysis. If $y_1,\ldots,y_{d+1}$ has a multivariate lognormal distribution and shape $(x_1,\ldots,x_d)$, defined by

$$x_i = y_i/(y_1 + \ldots + y_{d+1}) \quad (i = 1,\ldots,d),$$

is independent of *additive size* $z = y_1 + \ldots + y_{d+1}$ then the distribution of y is degenerate, being concentrated in a straight line in (d+1)-dimensional space. We shall see later, in §13 and Aitchison (24:1980), that this result, while correct, need not be a deterrent since a slight modification in modelling allows full exploitation of all the tractable features of lognormal distributions in this area of application.

5. Brown and Sanders (1980) provide a new insight into lognormal

genesis by their representation of a classification procedure in terms of the relative density of two mutually singular distributions. This is of particular interest since it confirms a conjecture of Aitchison and Brown (1:1957, p.27) that 'it is a curious fact that when a large number of items is classified on some homogeneity principle, the variate defined as the number of items in a class is often approximately log-normal.'

2.3 Some new applications

The combination of appropriateness and tractability continues to prove attractive to model-builders in an ever-increasing variety of disciplines. Any revision of Aitchison and Brown (1:1957) would involve substantial enlargement of the chapters on applications, where model selection leads naturally to lognormal distributions. Only a few of the more interesting additions can be singled out here.

1. In *economics*, Hart (1960) uses the measure of concentration analysis of Aitchison and Brown (1:1957, Chapter 11) in the study of business concentration; Cramer (1962) extends the cumulative lognormal Engel curve approach of Aitchison and Brown (1:1957) to the ownership of major consumer durables with the effects of income and wealth studied through the use of a bivariate lognormal model; Bain (1964) in his analysis of the growth of television ownership in the United Kingdom makes use of both cross-sectional and time series data through a cumulative lognormal growth model.
2. In the study of *language and language behaviour* Somers (1959), Rapoport (1964) and Carroll (1968) have all studied the effect-

iveness of lognormal models in word-frequency analysis.

3. In *climatology* Thom (1963) uses a bivariate lognormal model to describe the variability of path width and length of tornadoes, and exploits the multiplicative reproductive property in investigating tornado path areas.

4. In *palaeontology* MacGillavray (1965) uses lognormal modelling in relation to the variability of larger Foraminifera, and is particularly concerned with the choice of the threshhold of the three-parameter model. A similar form of application is by Schoener and Janzen (1968) in their study of environmental effects on the pattern of variability of insect size.

5. In *metallurgy* Schückher (1966) uses lognormal models in a wide-ranging study of one-, two- and three-dimensional grain size of metals.

6. For a recent strong advocacy of lognormal models in *chemical applications* see Siano (1972).

AITCHISON, J. and BROWN, J.A.C. (1957)

The Lognormal Distribution

Cambridge University Press

A copy of this book is reproduced in volume 2

3 GENERAL TECHNIQUES OF PARAMETRIC ESTIMATION AND HYPOTHESIS TESTING

3.1 Background

A popular and usually effective method of statistical estimation is the method of maximum likelihood. In the setting of §1 the likelihood function $L(\cdot|x)$ for given observation x in e is defined as the non-negative function on the q-dimensional parameter set Θ specified by

$$L(\theta|x) = p(x|\theta) \qquad (\theta \in \Theta).$$

A maximum-likelihood estimator is then defined as a function $\hat{\theta} : X \to \Theta$ such that

$$L\{\hat{\theta}(x)|x\} = \max_{\theta\in\Theta} L(\theta|x).$$

When an explicit solution is not available there are numerical methods of iterating towards a specific estimate $\hat{\theta}(x)$. Of these by far the most popular has been the Newton-Raphson method. For expository purposes here one particular version will be sufficient background to indicate the nature of the developments in methodology. Let D denote the q-vector derivative operator with ith component $\partial/\partial\theta_i$ and $D^2 = DD^T$ the $q \times q$ matrix second-derivative operator with (i,j)th component $\partial^2/\partial\theta_i\partial\theta_j$. Write

$$\ell(\theta|x) = D \log L(\theta|x),$$

$$B(\theta) = E_{p(x|\theta)}\{-D^2 \log L(\theta|x)\}.$$

Then the Newton-Raphson scheme for generating successive iterates

$\theta^{(r)}$ $(r = 0,1,\ldots)$ to $\hat{\theta}(x)$ is:

$$\theta^{(r)} = \theta^{(r-1)} + B^{-1}(\theta^{(r-1)})\ell(\theta^{(r-1)}|x) \quad (r = 1,2,\ldots)$$

with the value of B^{-1} at convergence providing an estimate of the covariance matrix of $\hat{\theta}$.

In the area of hypothesis testing a common technique for producing a test of a hypothesis ω $(\subset \Theta)$ against the alternative $\Theta - \omega$, or within the model Θ, is the generalised likelihood ratio criterion, using a critical region of the form

$$\{x : \Lambda(x) : \max_{\Theta} L(\theta|x)/\max_{\omega} L(\theta|x) > c\},$$

where the critical value c has to be determined to ensure the prescribed significance level. When ω is specified as the null space of some vector function h with r functionally independent components, so that

$$\omega = \{\theta \in \Theta : h(\theta) = 0\},$$

then c can be determined approximately 'for large sample sizes' as $\chi^2(r;1-\alpha)$, the $(1-\alpha)$ fractile of the chi-squared distribution with r degrees of freedom (Wilks, 1938).

The difficulty in applying such a test procedure is the need to determine not only the unrestricted maximum likelihood estimate $\hat{\theta}$ but also the restricted maximum likelihood estimate $\dot{\theta}$ which maximises L under the restriction $\theta \in \omega$. Wald (1943) had presented a method of avoiding this by devising an asymptotically equivalent test which required the evaluation only of $\hat{\theta}$. The test, which essentially asks if $\hat{\theta}$ approximately satisfies $h(\hat{\theta}) = 0$, is based on the critical region

$$[x : \{h^T(H^T B^{-1} H)^{-1} h\}_{\hat{\theta}} > \chi^2(r;1-\alpha)],$$

where H is the $q \times r$ matrix of first derivatives of the h functions with respect to the parameters, with (i,j)th component $\partial h_j(\theta)/\partial \theta_i$.

At this stage of the development of maximum likelihood estimation and these associated tests of significance there were a number of important questions remaining to be answered.

1. Suppose that the Wald test fails to reject ω. We shall then naturally wish to proceed with the estimation of θ^* under ω. For this restricted or constrained case is there any counterpart of the valuable Newton-Raphson method which will provide a systematic, reliable and easily programmed means of reaching $\dot{\theta}(x)$?
2. What if ω is most simply specified by freedom equations, setting $\theta = g(\alpha)$, where α is a vector of lower dimension than q? From this specification of ω as the range space of the vector function g it may prove difficult to arrive at the corresponding constraint equations $h(\theta) = 0$ and so apply the Wald test. In such circumstances it is clearly relatively straightforward to evaluate the maximum likelihood estimator of α, say $\dot{\alpha}$, and hence the restricted maximum likelihood estimator $\dot{\theta} = g(\dot{\alpha})$ of θ. We are thus led to the question: Just as the Wald test of ω requires only $\hat{\theta}$ is there a counterpart which requires only the evaluation of $\dot{\theta}$?
3. What if the natural specification of ω is in terms of a mixture of constraint and freedom equations? What aids to maximum likelihood estimation and hypothesis testing can be devised?

The answers to these questions were provided in Aitchison and Silvey (2:1958), Silvey (1959) and Aitchison and Silvey (3:1960), with later developments to aspects of multiple hypothesis testing in Aitchison (4:1962, 5:1964, 6:1965). The immediate motivation for these developments had arisen from an awareness by these authors that in their consultative work there was an increasing number of problems of this type, and that a general procedure that became automatic once the problem has been formulated could considerably ease the burden of the consulting statistician. For example, Aitchison had seen the need for such procedures in the simultaneous use of cross-section and time-series data in family budget analysis; see Aitchison and Brown (1:1957, Chapter 12) and Stone, Aitchison and Brown (1955), who indicate the need for attention to such problems. Also some aspects of a consultative problem of Silvey, reported in Aitchison and Silvey (1957) and concerned with a generalised probit-type model for the life cycle of a certain insect, would have been greatly simplified had such a theory of testing been available. Thus there was considerable incentive to obtain an easily applicable system.

The following is a summary of what was achieved in Aitchison and Silvey (2:1958), Silvey (1959) and Aitchison and Silvey (3:1960).

3.2 Contributions to estimation

Aitchison and Silvey (2:1958) concentrate on the problem of maximum likelihood estimation through the Lagrange-multiplier technique, maximising

$$\log L(\theta|x) + \lambda^T h(\theta)$$

with respect to (θ,λ), that is over $\Theta \times R^r$, where λ is an r-

dimensional vector of Lagrange multipliers. The results apply to the situation where e consists of n replicates of f.

1. Under certain carefully specified regularity conditions on the parametric form and on the constraint function there exists a solution of the first derivative equations

$$\ell(\theta|x) + H(\theta)\lambda = 0,$$

$$h(\theta) = 0,$$

and the solution is a consistent estimator of θ^*.

2. This solution provides not just a local maximum of L but an absolute maximum, and so a maximum likelihood estimator of θ^*.

3. From (1) and (2) the consistency property of the maximum likelihood estimator is established.

4. The asymptotic distribution of $\dot{\theta}$, $\hat{\lambda}$ is obtained: $\dot{\theta}$ and $\hat{\lambda}$ are asymptotically independent with normal distributions $N[\theta^*, \{B^{-1} - B^{-1}H(H^TB^{-1}H)^{-1}H^TB^{-1}\}_{\theta^*}]$ and $N[0, (H^TB^{-1}H)^{-1}_{\theta^*}]$, respectively, under the hypothesis ω.

5. A simple iterative procedure similar to the Newton-Raphson method is available for the determination of a sequence of iterates $\theta^{(r)}$, $\lambda^{(r)}$ leading, with convergence, to $\dot{\theta}$, $\hat{\lambda}$:

$$\begin{bmatrix} \theta^{(r)} \\ \lambda^{(r)} \end{bmatrix} = \begin{bmatrix} \theta^{(r-1)} \\ 0 \end{bmatrix} + \begin{bmatrix} B & -H \\ -H^T & 0 \end{bmatrix}^{-1}_{\theta^{(r-1)}} \begin{bmatrix} \ell(\theta^{(r-1)}|x) \\ h(\theta^{(r-1)}) \end{bmatrix}.$$

6. Finally, the paper indicates that as an alternative to the generalised likelihood ratio test and the Wald test of ω within Θ the Lagrangian multiplier test statistic

$$\hat{\lambda}(H^TB^{-1}H)_{\dot{\theta}}\hat{\lambda} = \ell^T(\dot{\theta}|x)B^{-1}_{\dot{\theta}}\ell(\dot{\theta}|x)$$

may be used against, asymptotically, $\chi^2(r;1-\alpha)$.

3.3 Contributions to hypothesis-testing

No attempt is made to justify the use of the Lagrangian multiplier test in Aitchison and Silvey (2:1958). The asymptotic equivalence of

(i) the generalised likelihood ratio test,

(ii) the Wald test,

(iii) the Lagrangian multiplier test,

is established by Silvey (1959), who also considers the adjustments necessary for the important case where the information matrix B is singular. This singularity reflects the presence of unidentifiability of parameters, requiring the introduction of identifiability restrictions as part of the constraint equations $h(\theta) = 0$. Suppose that $h = \{h_1, h_2\}$ with h_1 the constraint function for identifiability and h_2 the constraint function defining the hypothesis. Then Silvey (1959) shows that all that is required is to replace the matrix B by $B + H_1 H_1^T$, where $H = [H_1, H_2]$ with an obvious partitioning, and to replace r by $r - r_1$ where r_1 is the dimension of vector function h_1.

These two papers, Aitchison and Silvey (2:1958) and Silvey (1959), made available easy-to-use practical tools for the consulting statistician faced with a parametric estimation or hypothesis testing situation. All that is required is to identify the likelihood L and the constraint function h. Once this formulation is complete the rest is straightforward technical mathematics (some differentiation) and computing. Now Aitchison and Silvey (3:1960) set out the practical application of the methods, discussing carefully the relative merits of the different tests in relation to the practical situation. In addition to its expository and

illustrative role this paper provided three new aspects.

1. Identification in any multinomial situation of the Lagrange-multiplier test statistic with the familiar chi-squared goodness-of-fit statistic $\Sigma(\text{observed-expected})^2/\text{expected}$.
2. A discussion of the case where the constraints are specified in terms of freedom equations.
3. A method of dealing with a mixed specification, some constraint and some freedom equations.

AITCHISON, J. and SILVEY, S.D. (1958)

Maximum-likelihood estimation of parameters subject to restraints

Reprinted from *Ann. Math. Statist.* 29, 813-28

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 29, No. 3, September, 1958
Printed in U.S.A.

MAXIMUM-LIKELIHOOD ESTIMATION OF PARAMETERS SUBJECT TO RESTRAINTS

BY J. AITCHISON AND S. D. SILVEY

University of Glasgow

Summary. The estimation of a parameter lying in a subset of a set of possibl parameters is considered. This subset is the null space of a well-behaved functio and the estimator considered lies in the subset and is a solution of likelihoo equations containing a Lagrangian multiplier. It is proved that, under certai conditions analogous to those of Cramér, these equations have a solution whic gives a local maximum of the likelihood function. The asymptotic distribution c this 'restricted maximum likelihood estimator' and an iterative method of solv ing the equations are discussed. Finally a test is introduced of the hypothesi that the true parameter does lie in the subset; this test, which is of wide appli cability, makes use of the distribution of the random Lagrangian multiplie appearing in the likelihood equations.

1. Introduction. Quite frequently in statistical theory the natural way o building up a mathematical model of an experiment leads to the description o the experiment by a random variable X whose distribution function F depend: on s parameters $\theta_1, \theta_2, \cdots, \theta_s$, which are not mathematically independen but satisfy r functional relationships $h_i(\theta_1, \theta_2, \cdots, \theta_s) = 0$, $i = 1, 2, \cdots, r$ $r < s$. In many cases where such a natural description arises it is possible t solve the r equations $h_i(\theta_1, \theta_2, \cdots, \theta_s) = 0$ for r of the parameters in terms o the remaining $s - r$, to express the distribution function F in terms of thes remaining parameters only and, given observations on X, to estimate thes $s - r$ unrestricted parameters by the method of maximum likelihood. Thi procedure has two disadvantages. First, it may be impossible to express r of th parameters explicitly in terms of the remaining $s - r$ and second, interest may lie in estimating all of the parameters simultaneously, in which case a symmetrical procedure for so doing is certainly desirable. The natural symmetric method for maximum-likelihood estimation in this case is achieved by the introduction of Lagrangian multipliers and it is this method that we will consider in this paper.

2. Formulation of the problem. In this section we will formulate more precisely the problem to be considered.

We will denote m-dimensional Euclidian space by $\mathfrak{R}^m$, $m = 1, 2, 3, \cdots$. A point in $\mathfrak{R}^s$, denoted by $\theta = (\theta_1, \theta_2, \cdots, \theta_s)$ will represent a value of a parameter. There is a particular point $\theta_0 = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_s^{(0)})$ in $\mathfrak{R}^s$ which is the true, though unknown, parameter value. Corresponding to each θ in some neighbour-

Received February 18, 1957.

hood of θ_0, say in $U_\alpha = \{\theta : \| \theta - \theta_0 \| \leqq \alpha\}$, is a probability density function f_θ defined on $\mathfrak{R}^1$ and we will denote the value of f_θ at the point $t \in \mathfrak{R}^1$ by $f(t, \theta)$. The probability density function f_{θ_0} defines a probability measure on $\mathfrak{R}^1$ and we will assume that, with respect to this measure, for almost all t, the partial derivatives $\partial \log f(t, \theta)/\partial\theta_i$, $i = 1, 2, \cdots, s$, exist for every θ in U_α.

There is given a continuous function h from $\mathfrak{R}^s$ into $\mathfrak{R}^r$, $r < s$, defined by $h(\theta) = (h_1(\theta), h_2(\theta), \cdots, h_r(\theta))$, which is such that, for every θ in U_α, the partial derivatives $\partial h_j(\theta)/\partial\theta_i$, $i = 1, 2, \cdots, s, j = 1, 2, \cdots, r$, exist. The function h has the further property that $h(\theta_0) = 0$.

A point in $\mathfrak{R}^n$ denoted by $x = (x_1, x_2, \cdots, x_n)$ will be regarded as representing a set of n independent observations on a random variable whose probability density function is f_{θ_0} and we use the fact that points in $\mathfrak{R}^n$ are being so regarded to define, in the usual way, a probability measure on $\mathfrak{R}^n$, for each n. Subsequent statements regarding the probabilities of sets in $\mathfrak{R}^n$ will refer to this particular probability measure.

It will be convenient to use also matrix representation for points in $\mathfrak{R}^m$ and for linear operators from one Euclidian space to another and we will use the convention that, for example, θ is the $s \times 1$ column vector representing the point θ in $\mathfrak{R}^s$, and H, an $s \times r$ matrix, represents a linear operator H from $\mathfrak{R}^r$ into $\mathfrak{R}^s$.

The log-likelihood function L is defined on a subset of $\mathfrak{R}^n \times \mathfrak{R}^s$ by

$$L(x, \theta) = \sum_{i=1}^{n} \log f(x_i, \theta).$$

If H_θ denotes the $s \times r$ matrix $(\partial h_j(\theta)/\partial\theta_i)$, and if λ is a Lagrangian multiplier in $\mathfrak{R}^r$, then we propose to estimate the unknown parameter θ_0 by a solution, if such exists, of the equations

$$\ell(x, \theta) + H_\theta \lambda = 0 \tag{2.1}$$

$$h(\theta) = 0, \tag{2.2}$$

where $\ell(x, \theta)$ is the point in $\mathfrak{R}^s$ whose ith component is $\partial L(x, \theta)/\partial\theta_i$.

We will show that, under certain fairly general conditions, if x belongs to a set whose probability measure tends to 1 as $n \to \infty$, these equations have a solution $\hat{\theta}(x)$, $\hat{\lambda}(x)$, where $\hat{\theta}(x)$ is near θ_0 and $\hat{\theta}(x)$ maximises $L(x, \theta)$ subject to the condition $h(\theta) = 0$. The definition of $\hat{\theta}$ and $\hat{\lambda}$ will then be extended in a natural way to the whole of $\mathfrak{R}^n$ and we will show that the random variables thus defined are asymptotically jointly normally distributed. We will then consider an iterative procedure for solving the equations (2.1) and (2.2). Finally tests of the adequacy of the model will be introduced.

3. Existence of a solution. The proof that we will give of the existence of a solution of the equations (2.1) and (2.2) is based on the same principle as a proof given by Cramér [2] of the existence of a maximum likelihood estimate of a parameter in $\mathfrak{R}^1$. However the presence of the restraining condition $h(\theta) = 0$ in the situation we are discussing makes our proof more intricate in detail than a

straightforward generalisation of Cramér's proof to a parameter in $\mathfrak{R}^s$ would be. And we start by indicating the main lines of the proof.

We set out to show that, under certain conditions, if δ is a sufficiently small given number and if n is sufficiently large, then, for a set of x whose probability measure is near 1, the equations (2.1) and (2.2) have a solution $\hat{\theta}(x)$, $\hat{\lambda}(x)$, where $\hat{\theta}(x) \in U_\delta$. We will demand that in U_α the function $\log f(x, \cdot)$ should possess partial derivatives of the third order and the components of the function h should possess partial derivatives of the second order. Then it will be possible, by expanding the components of $\ell(x, \theta)$ and $h(\theta)$ about θ_0 to express the equations (in matrix notation) in the form

$$l(x, \theta_0) + M_{x, \theta_0}(\theta - \theta_0) + v^{(1)}(x, \theta) + H_\theta \lambda = 0, \tag{3.1}$$

$$H'_{\theta_0}(\theta - \theta_0) + v^{(2)}(\theta) = 0, \tag{3.2}$$

where

(i) M_{x, θ_0} is the matrix $(\partial^2 L(x, \theta_0)/\partial\theta_i \partial\theta_j)$,

(ii) $v^{(1)}(x, \theta)$ is a vector whose mth component may be expressed in the form $\frac{1}{2}(\theta - \theta_0)' L_m (\theta - \theta_0)$, L_m being the matrix $(\partial^3 L(x, \theta^{(m,1)})/\partial\theta_m \partial\theta_i \partial\theta_j)$, $i, j = 1, 2, \cdots, s$, and $\theta^{(m,1)}$ a point such that $\| \theta^{(m,1)} - \theta_0 \| < \| \theta - \theta_0 \|$.

(iii) $v^{(2)}(\theta)$ is a vector whose mth component is $\frac{1}{2}(\theta - \theta_0)' H_m (\theta - \theta_0)$, H_m being the matrix $(\partial^2 h_m(\theta^{(m,2)})/\partial\theta_i \partial\theta_j)$, $i, j = 1, 2, \cdots, s$, and $\theta^{(m,2)}$ a point such that $\| \theta^{(m,2)} - \theta_0 \| < \| \theta - \theta_0 \|$.

Further conditions imposed on f, which are almost a straightforward generalisation of Cramér's conditions [2], will ensure that, for large enough n, there is a set of x whose probability measure is near 1 such that, if x belongs to this set,

(i) $\| (1/n)\ell(x, \theta_0) \|$ is small,

(ii) $-(1/n)M_{x, \theta_0}$ is near a certain positive definite matrix B_{θ_0} and

(iii) the elements of $(1/n)L_m$ are bounded for $\theta \in U_\delta$. By dividing (3.1) by n we will then be able to express this equation in the form

$$-B_{\theta_0}(\theta - \theta_0) + \frac{1}{n} H_\theta \lambda + \delta^2 v^{(3)}(x, \theta) = 0 \tag{3.3}$$

where $\| v^{(3)}(x, \theta) \|$ is bounded for $\theta \in U_\delta$. In addition we will demand that, for $\theta \in U_\alpha$, the second order derivatives of the components of h should be bounded. Then we will be able to express (3.2) in the form

$$H'_{\theta_0}(\theta - \theta_0) + \delta^2 v^{(4)}(\theta) = 0 \tag{3.4}$$

where $\| v^{(4)}(\theta) \|$ is bounded for $\theta \in U_\delta$.

If the equations (3.3) and (3.4) have a solution, then by pre-multiplying (3.3) by $H'_{\theta_0} B^{-1}_{\theta_0}$ and substituting for $H'_{\theta_0}(\theta - \theta_0)$ from (3.4) we find that the values of θ and λ satisfying these equations also satisfy an equation of the form

$$H'_{\theta_0} B^{-1}_{\theta_0} H_\theta \left(\frac{1}{n}\lambda\right) + \delta^2 v^{(5)}(x, \theta) = 0. \tag{3.5}$$

We will impose conditions on h which ensure that the matrix $H'_{\theta_0} B^{-1}_{\theta_0} H_\theta$ is non-

singular and the elements of its inverse are bounded functions of θ for $\theta \in U_\delta$. Then it will be possible to solve equation (3.5) for λ in terms of θ and on substitution in (3.3) we will obtain the result that any value of θ in U_δ for which equations (3.3) and (3.4) are satisfied is also a solution of an equation of the form

$$-B_{\theta_0}(\theta - \theta_0) + \delta^2 v(x, \theta) = 0 \tag{3.6}$$

where $\| v(x, \theta) \|$ is bounded for $\theta \in U_\delta$.

Conversely it will be shown that if the equation (3.6) has a solution $\hat{\theta}(x) \in U$ then $\hat{\theta}(x)$ leads to a solution $\hat{\theta}(x)$, $\hat{\lambda}(x)$ of equations (2.1) and (2.2). We will then use the fact that B_{θ_0} is a positive definite matrix to prove that, if δ is sufficiently small, (3.6) has a solution in U_δ.

This outline of the method of proof to be adopted provides the motivation for the introduction of conditions on f and h which we now discuss.

Conditions on f. The following conditions on the function f appear complicated and restrictive from the mathematical point of view. In fact they will be satisfied in most practical estimation problems.

$\mathfrak{F}1$. *For every $\theta \in U_\alpha$ and for almost all $t \in \mathfrak{R}^1$ (almost all with respect to the probability measure on $\mathfrak{R}^1$ defined by f_{θ_0}), the derivatives*

$$\frac{\partial \log f(t, \theta)}{\partial \theta_i}, \quad \frac{\partial^2 \log f(t, \theta)}{\partial \theta_i \partial \theta_j} \quad \text{and} \quad \frac{\partial^3 \log f(t, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}, \qquad i, j, k = 1, 2, \cdots, s$$

exist, and the first and second order derivatives are continuous functions of θ.

$\mathfrak{F}2$. *For every $\theta \in U_\alpha$ and for $i, j = 1, 2, \cdots, s$, $|\partial f(t, \theta)/\partial \theta_i| < F_1(t)$ and $|\partial^2 f(t, \theta)/\partial \theta_i \partial \theta_j| < F_2(t)$, where F_1 and F_2 are finitely integrable over $(-\infty, \infty)$.*

$\mathfrak{F}3$. *For every $\theta \in U_\alpha$ and $i, j, k = 1, 2, \cdots, s$, $|\partial^3 \log f(t, \theta)/\partial \theta_i \partial \theta_j \partial \theta_k| < F_3(t)$, where $\int_{-\infty}^{\infty} F_3(t) f(t, \theta_0)\, dt$ is finite and equal to κ_1, say.*

$\mathfrak{F}4$.
$$b_{ij} = \int_{-\infty}^{\infty} \frac{\partial \log f(t, \theta_0)}{\partial \theta_i} \frac{\partial \log f(t, \theta_0)}{\partial \theta_j} f(t, \theta_0)\, dt$$

is finite for $i, j = 1, 2, \cdots, s$, and the matrix $B_{\theta_0} = (b_{ij})$ is positive definite with minimum latent root μ_0.

The conditions $\mathfrak{F}3$ and $\mathfrak{F}4$ are apparently less stringent than a straightforward generalisation of Cramér's corresponding conditions would be. In §6 we return to this point.

If f satisfies these conditions then for any given positive numbers $\delta < \alpha$ and $\epsilon < 1$ and for sufficiently large n, say $n \geqq n(\delta, \epsilon)$, there exists a set $X_n \subset \mathfrak{R}$ with the properties

$\mathfrak{X}1$. $\Pr\{X_n\} > 1 - \epsilon$.

$\mathfrak{X}2$. $\left\| \frac{1}{n} \ell(x, \theta_0) \right\| < \delta^2$, if $x \in X_n$.

$\mathfrak{X}3$. $\frac{1}{n} M_{x, \theta_0}$ can be expressed in the form $-B_{\theta_0} + \delta m_{x, \theta_0}$,

where m_{x,θ_0} is an $s \times s$ matrix the moduli of whose elements are bounded by 1, if $x \in X_n$.

$\mathfrak{X}4$. For *every* $\theta \in U_\alpha$ and $i, j, k = 1, 2, \cdots, s$,

$$\left|\frac{1}{n}\frac{\partial^3 L(x,\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| < 2\kappa_1$$

if $x \in X_n$.

The proof of these results is similar to the proof of the corresponding results given by Cramér [2] in the case of a parameter in $\mathfrak{R}^1$ and we merely remark that the conditions $\mathfrak{F}1$–4 imply (as they are designed to imply) that

(i) $(1/n)l(\cdot, \theta_0)$ converges in probability to $0 \in \mathfrak{R}^s$,

(ii) $(1/n)M_{\cdot,\theta_0}$ converges in probability to $-B_{\theta_0}$, and

(iii) if $G(x) = 1/n\sum_{i=1}^n F_3(x_i)$, then the random variable G converges in probability to κ_1 and

$$\frac{1}{n}\left|\frac{\partial^3 L(x,\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| = \frac{1}{n}\left|\sum_{l=1}^n \frac{\partial^3 \log f(x_l,\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| < G(x),$$

by $\mathfrak{F}3$.

In future when we refer to a set X_n we imply that n is sufficiently large for the existence of a set in $\mathfrak{R}^n$ with the properties $\mathfrak{X}1$–4 and that the set X_n referred to has these properties.

As has already been indicated, one of the main purposes of the introduction of the conditions $\mathfrak{F}$ was to ensure that (3.1) could be expressed in the form (3.3). Now if the conditions $\mathfrak{F}$ are satisfied, if $x \in X_n$ and $\theta \in U_\delta$, it is easily verified that

(i) by $\mathfrak{X}2$, $(1/n\delta^2)\| l(x,\theta_0)\| < 1$,

(ii) by $\mathfrak{X}3$, $(1/\delta)\| m_{x,\theta_0}(\theta-\theta_0)\| \leqq s^2$,

(iii) by $\mathfrak{X}4$, $(1/n\delta^2)\| v^{(1)}(x,\theta)\| < (1/\delta^2)s^3\kappa_1\| \theta-\theta_0\|^2 \leqq s^3\kappa_1$.

It follows that (3.1) can then be expressed in the form (3.3) and

$$\| v^{(3)}(x,\theta)\| < 1 + s^2 + s^3\kappa_1, \quad \text{when} \quad x \in X_n \quad \text{and} \quad \theta \in U_\delta.$$

Conditions on h. We impose the following conditions on the function h.

$\mathfrak{H}1$. *For every* $\theta \in U_\alpha$ *the partial derivatives* $\partial h_k(\theta)/\partial\theta_i$, $i = 1, 2, \cdots, s$, $k = 1, 2, \cdots, r$, *exist and these are continuous functions of* θ.

$\mathfrak{H}2$. *For every* $\theta \in U_\alpha$ *the partial derivatives* $\partial^2 h_k(\theta)/\partial\theta_i\partial\theta_j$, $i, j = 1, 2, \cdots, s$, $k = 1, 2, \cdots, r$, *exist and* $|\partial^2 h_k(\theta)/\partial\theta_i\partial\theta_j| < 2\kappa_2$, *a given constant, for all* i, j *and* k.

$\mathfrak{H}3$. *The* $s \times r$ *matrix* H_{θ_0} *is of rank* r.

The condition $\mathfrak{H}2$ is introduced to ensure that when (3.2) is expressed in the form (3.4), $\| v^{(4)}(\theta)\|$ is bounded for $\theta \in U_\delta$. It is clear that it does ensure this since, as is easily verified, by $\mathfrak{H}2$, $\| v^{(2)}(\theta)\| < s^3\kappa_2\| \theta-\theta_0\|^2$ and so $\| v^{(4)}(\theta)\| = (1/\delta^2)\| v^{(2)}(\theta)\| < s^3\kappa_2$ if $\theta \in U_\delta$.

Also the condition $\mathfrak{C}3$ implies that the matrix $H'_{\theta_0}B^{-1}_{\theta_0}H_{\theta_0}$ is positive definite, since the matrix $B^{-1}_{\theta_0}$ is positive definite. Since the elements of H_θ are, by $\mathfrak{C}1$, continuous functions of θ it follows that there exists a neighbourhood of θ_0 in which $\det (H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)$ is bounded away from zero, and we may assume that this neighbourhood is U_α. (This assumption merely involves choosing α small enough initially). This means that when $\theta \in U_\alpha$ we can solve the equation (3.5) for λ in terms of θ. Furthermore the elements of the matrix $(H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)^{-1}$ are then continuous functions on U_α since the elements of H_θ are continuous and $\det (H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)$ is bounded away from zero. Since U_α is a closed set it follows that the elements of $(H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)^{-1}$ are uniformly bounded on U_α. This result, together with the results that $\| v^{(3)}(x, \theta) \|$ and $\| v^{(4)}(\theta) \|$ are bounded on U_δ, enable us to prove that when λ is eliminated from (3.3) and (3.4), and (3.6) is obtained, then in (3.6) $\| v(x, \theta) \|$ is bounded on U_δ, if $x \in X_n$.

We have now gone a considerable way towards proving the main part of the following lemma.

LEMMA 1. *Subject to the conditions $\mathfrak{F}$ and $\mathfrak{C}$, if $\delta < \alpha$ and $\epsilon < 1$ are given positive numbers and if $x \in X_n$, then the equations* (2.1) *and* (2.2) *have a solution $\hat\theta(x)$, $\hat\lambda(x)$ such that $\hat\theta(x) \in U_\delta$, if and only if $\hat\theta(x)$ satisfies a certain equation of the form $-B_{\theta_0}(\theta - \theta_0) + \delta^2 v(x, \theta) = 0$. In this equation $v(x, \cdot)$ is a continuous function on U_δ and $\| v(x, \theta) \|$ is bounded for $\theta \in U_\delta$ by a positive number κ_3, say.*

PROOF. The fact that the condition is necessary has virtually been established already. On eliminating λ from (2.1) and (2.2) by the method outlined at the beginning of §3 we obtain, in matrix notation, the following explicit expression for (3.6)

$$-B_{\theta_0}(\theta - \theta_0) - H_\theta(H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)^{-1}\{v^{(2)}(\theta) + H'_{\theta_0}B^{-1}_{\theta_0}v^{(6)}(x, \theta)\} + v^{(6)}(x, \theta) = 0, \tag{3.7}$$

where

$$v^{(6)}(x, \theta) = \delta^2 v^{(3)}(x, \theta) = \frac{1}{n} l(x, \theta) + B_{\theta_0}(\theta - \theta_0), \tag{3.8}$$

and

$$v^{(2)}(\theta) = \delta^2 v^{(4)}(\theta) = h(\theta) - H'_{\theta_0}(\theta - \theta_0). \tag{3.9}$$

Hence in (3.6),

$$v(x, \theta) = -H_\theta(H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)^{-1}\{v^{(4)}(\theta) + H'_{\theta_0}B^{-1}_{\theta_0}v^{(3)}(x, \theta)\} + v^{(3)}(x, \theta). \tag{3.10}$$

The fact that $v(x, \cdot)$ is a continuous function on U_δ and that $\| v(x, \theta) \|$ is bounded for $\theta \in U_\delta$ follows from (3.8), (3.9) and (3.10), in virtue of the discussion of $v^{(3)}(x, \theta)$, $v^{(4)}(\theta)$ and $(H'_{\theta_0}B^{-1}_{\theta_0}H_\theta)^{-1}$ above.

Turning to the sufficiency of the condition we now suppose that the equation

(3.7) has a root $\hat{\theta}(x) \in U_\delta$. Then, writing $\hat{\theta}$ instead of $\hat{\theta}(x)$ for brevity, we obtain on premultiplication of (3.7) by $H'_{\theta_0}B^{-1}_{\theta_0}$,

$$-H'_{\theta_0}(\hat{\theta} - \theta_0) - v^{(2)}(\hat{\theta}) = 0, \tag{3.11}$$

i.e., by (3.9),

$$h(\hat{\theta}) = 0.$$

Substitution for $v^{(2)}(\hat{\theta})$ from (3.11) and for $v^{(6)}(x, \hat{\theta})$ from (3.8), in (3.7) gives

$$l(x, \hat{\theta}) = H_{\hat{\theta}}(H'_{\theta_0}B^{-1}_{\theta_0}H_{\hat{\theta}})^{-1}H'_{\theta_0}B^{-1}_{\theta_0}l(x, \hat{\theta}),$$

or, if we write $Q_{\hat{\theta}}$ for $(H'_{\theta_0}B^{-1}_{\theta_0}H_{\hat{\theta}})^{-1}H'_{\theta_0}B^{-1}_{\theta_0}$,

$$l(x, \hat{\theta}) = H_{\hat{\theta}}Q_{\hat{\theta}}l(x, \hat{\theta}). \tag{3.12}$$

If we now define $\hat{\lambda}(x)$ by

$$\hat{\lambda}(x) = -Q_{\hat{\theta}}l(x, \hat{\theta}),$$

then

$$l(x, \hat{\theta}) + H_{\hat{\theta}}\hat{\lambda}(x) = 0,$$

and $\hat{\theta}(x)$, $\hat{\lambda}(x)$ satisfy the equations (2.1) and (2.2).

In order to prove that the equation (3.6) has a root in U_δ, if δ is sufficiently small, we will require the following lemma.

LEMMA 2: *If g is a continuous function mapping $\mathcal{R}^r$ into itself with the property that, for every θ such that $\| \theta \| = 1$, $\theta' g(\theta) < 0$, then there exists a point $\hat{\theta}$ such that $\| \hat{\theta} \| < 1$ and $g(\hat{\theta}) = 0$.*

PROOF. For the proof of this result we are indebted to Mr. J. M. Michael who has proved that this result is equivalent to Brouwer's fixed point theorem [4]. A direct proof from the latter theorem is as follows.

We suppose that $g(\theta) \neq 0$ for any θ such that $\| \theta \| \leqq 1$. Then the function g_1, defined on the unit sphere in $\mathcal{R}^r$ by

$$g_1(\theta) = \frac{g(\theta)}{\| g(\theta) \|},$$

is a continuous function mapping this unit sphere into itself. Hence by the fixed point theorem there is a point θ^* in the unit sphere such that $\theta^* = g_1(\theta^*)$. Also since $\| g_1(\theta) \| = 1$ for every θ in the unit sphere, it follows that $\| \theta^* \| = 1$, and $\theta^{*\prime} g_1(\theta^*) = \theta^{*\prime}\theta^* = 1 > 0$. But this contradicts the fact that $\theta' g(\theta) < 0$ (and consequently that $\theta' g_1(\theta) < 0$) for every θ such that $\| \theta \| = 1$.

Hence there is a point $\hat{\theta}$ in the unit sphere such that $g(\hat{\theta}) = 0$. It is obvious that $\| \hat{\theta} \| \neq 1$. Hence $\| \hat{\theta} \| < 1$.

We are now in a position to prove the following existence theorem.

THEOREM 1. *Subject to the conditions $\mathfrak{F}$ and $\mathfrak{K}$, if δ is a sufficiently small given*

positive number, ϵ is a given positive number less than 1 and if $x \in X_n$, then the equations (2.1) *and* (2.2) *have a solution $\hat{\theta}(x)$, $\hat{\lambda}(x)$ such that $\hat{\theta}(x) \in U_\delta$.*

PROOF. We suppose $\delta < \alpha$ and $x \in X_n$. We consider (3.6) and define a function g on the unit sphere in $\mathfrak{R}^r$ by

$$g\left(\frac{\theta - \theta_0}{\delta}\right) = -B_{\theta_0}(\theta - \theta_0) + \delta^2 v(x, \theta).$$

By Lemma 1, $v(x, \cdot)$ is a continuous function on U_δ. Hence g is a continuous function on the unit sphere in $\mathfrak{R}^r$. Also

$$\frac{1}{\delta}(\theta - \theta_0)'g\left(\frac{\theta - \theta_0}{\delta}\right) = -\frac{1}{\delta}(\theta - \theta_0)'B_{\theta_0}(\theta - \theta_0) + \delta(\theta - \theta_0)'v(x, \theta)$$

$$\leqq -\frac{1}{\delta}\mu_0 \| \theta - \theta_0 \|^2 + \delta\kappa_3 \| \theta - \theta_0 \|,$$

if $\theta \in U_\delta$, since B_{θ_0} is positive definite with minimum latent root μ_0 and, by Lemma 1, $\| v(x, \theta) \| < \kappa_3$ when $\theta \in U_\delta$. Hence for every θ such that $\| \theta - \theta_0 \| = \delta$, we have

$$\frac{1}{\delta}(\theta - \theta_0)'g\left(\frac{\theta - \theta_0}{\delta}\right) \leqq \delta(\delta\kappa_3 - \mu_0)$$

$$< 0, \quad \text{if} \quad \delta < \frac{\mu_0}{\kappa_3}.$$

Hence if $\delta < \mu_0/\kappa_3$, it follows by Lemma 2 that there exists a point $\hat{\theta}(x)$ such that $\hat{\theta}(x) \in U_\delta$ and $g((\hat{\theta}(x) - \theta_0) / \delta) = 0$, i.e., $\hat{\theta}(x)$ is a solution of (3.6). The result follows by application of Lemma 1.

4. Existence of a maximum of $L(x, \theta)$. In this paragraph we will show that for sufficiently small δ, if $x \in X_n$, any solution of (3.6) in U_δ maximises $L(x, \theta)$ subject to the condition $h(\theta) = 0$.

We suppose that $x \in X_n$, that δ is small enough for Theorem 1 to apply and that $\hat{\theta}(x)$, written $\hat{\theta}$ for typographical brevity, is a solution in U_δ of (3.6). We let θ be a point in a neighbourhood of $\hat{\theta}$ contained in U_δ, such that $h(\theta) = 0$. (Such a neighbourhood exists since $\hat{\theta}$ is an interior point of U_δ.) Then by expanding $L(x, \theta)$ about $\hat{\theta}$ we have

$$L(x, \theta) - L(x, \hat{\theta}) = l'(x, \hat{\theta})(\theta - \hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})'M_{x,\theta^*}(\theta - \hat{\theta}) \tag{4.1}$$

where $M_{x,\theta^*} = (\partial^2 L(x, \theta^*) / \partial\theta_i \partial\theta_j)$ and $\theta^* \in U_\delta$.

We now consider separately the two terms in the right hand side of (4.1). By (3.12)

$$l'(x, \hat{\theta})(\theta - \hat{\theta}) = l'(x, \hat{\theta})Q_\delta' H_\delta'(\theta - \hat{\theta}).$$

Now

$$0 = h(\theta) - h(\hat{\theta}) = H_\delta'(\theta - \hat{\theta}) + r(\theta),$$

where, because of 3C2, by the same argument as was applied to $v^{(2)}(\theta)$ in (3.2),

(4.2) $$\| r(\theta) \| < s^3 \kappa_2 \| \theta - \hat{\theta} \|^2.$$

Hence

(4.3) $$l'(x, \hat{\theta})(\theta - \hat{\theta}) = -[Q_{\hat{\theta}} l(x, \hat{\theta})]' r(\theta).$$

By (3.8)

$$\frac{1}{n} l(x, \hat{\theta}) = -B_{\theta_0}(\hat{\theta} - \theta_0) + v^{(6)}(x, \hat{\theta}),$$

and so

$$\frac{1}{n} \| \ell(x, \hat{\theta}) \| < \kappa_4 \delta + \kappa_5 \delta^2, \quad \text{since} \quad \hat{\theta} \in U_\delta,$$

where κ_4 is a positive number depending only on the elements of B_{θ_0}, and, as above, $\kappa_5 = 1 + s^2 + s^3 \kappa_1$. Also the elements of $Q_{\hat{\theta}}$ are bounded by a number independent of δ, since $\hat{\theta} \in U_\alpha$. Hence

(4.4) $$\frac{1}{n} \| Q_\theta l(x, \hat{\theta}) \| < \kappa_6 \delta + \kappa_7 \delta^2,$$

where κ_6, κ_7 are positive numbers independent of δ. From (4.2), (4.3) and (4.4) it follows that

(4.5) $$\frac{1}{n} | l'(x, \hat{\theta})(\theta - \hat{\theta}) | < (\kappa_6 \delta + \kappa_7 \delta^2) s^3 \kappa_2 \| \theta - \hat{\theta} \|^2.$$

We now consider the second term of (4.1). By expanding the elements of M_{x,θ^*} about θ_0 we find that

$$\frac{1}{n} M_{x,\theta^*} = \frac{1}{n} M_{x,\theta_0} + m^*_{x,\theta^*},$$

where, as is easily shown using §4, the moduli of the elements of the matrix m^*_{x,θ^*} are less than $2s\kappa_1\delta$. Also by §3,

$$\frac{1}{n} M_{x,\theta_0} = -B_{\theta_0} + \delta m_{x,\theta_0},$$

and so

$$\frac{1}{n} M_{x,\theta^*} = -B_{\theta_0} + \delta m,$$

say, where m is a matrix whose elements are bounded by a number independent of δ. Hence

(4.6) $$\frac{1}{2n}(\theta - \hat{\theta})' M_{x,\theta^*}(\theta - \hat{\theta}) = -\frac{1}{2}(\theta - \hat{\theta})' B_{\theta_0}(\theta - \hat{\theta})$$

$$+ \frac{1}{2}\delta(\theta - \hat{\theta})' m(\theta - \hat{\theta}) < -\frac{1}{2}\mu_0 \| \theta - \hat{\theta} \|^2 + \kappa_8 \delta \| \theta - \hat{\theta} \|^2,$$

since B_{θ_0} is positive definite with minimum latent root μ_0, and the elements of m are bounded. Here κ_8 is a positive number depending only on the elements of m. Using (4.5) and (4.6) in (4.1) we find that there exist positive numbers κ_9, κ_{10}, independent of δ, such that

$$\frac{1}{n}[L(x,\theta) - L(x,\hat{\theta})] < \left(-\frac{1}{2}\mu_0 + \kappa_9\delta + \kappa_{10}\delta^2\right) \| \theta - \hat{\theta} \|^2.$$

It follows that if δ is sufficiently small then $L(x, \theta) < L(x, \hat{\theta})$, i.e., $L(x, \hat{\theta})$ is a maximum value of $L(x, \theta)$ subject to $h(\theta) = 0$.

We have thus established the fact that, if the conditions $\mathfrak{F}$ and $\mathfrak{K}$ are satisfied, there exists a consistent maximum likelihood estimator $\hat{\theta}$ of θ_0 satisfying the condition $h(\hat{\theta}) = 0$.

5. Asymptotic distributions. We return now to consideration of (3.1) and (3.2). We suppose that $x \in X_n$ and that $\hat{\theta}(x)$, $\hat{\lambda}(x)$ is a solution of these equations with $\hat{\theta}(x) \in U_\delta$, δ being small enough for such a solution to exist. Then, considering the equations from a slightly different viewpoint we have,

$$\frac{1}{n}1(x,\theta_0) - [B_{\theta_0} + \hat{b}(x)][\hat{\theta}(x) - \theta_0] + [H_{\theta_0} + \hat{h}(x)]\frac{1}{n}\hat{\lambda}(x) = 0, \tag{5.1}$$

$$[H'_{\theta_0} + \hat{h}^*(x)][\hat{\theta}(x) - \theta_0] = 0, \tag{5.2}$$

where $\hat{b}(x)$, $\hat{h}(x)$ and $\hat{h}^*(x)$ are matrices whose elements tend to 0 as δ (and hence $\| \hat{\theta}(x) - \theta_0 \|$) $\to 0$. We now prove the following lemma.

LEMMA 3. *The partitioned matrix*

$$\begin{bmatrix} B_{\theta_0} & -H_{\theta_0} \\ -H'_{\theta_0} & 0 \end{bmatrix}$$

is non-singular.

PROOF. For brevity we omit the suffix θ_0. Then we wish to find a matrix

$$\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix}$$

such that, in the usual notation,

$$\begin{bmatrix} B & -H \\ -H' & 0 \end{bmatrix}\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} = \begin{bmatrix} I_s & 0 \\ 0 & I_r \end{bmatrix}$$

and this requires

$$BP - HQ' = I_s, \tag{5.3}$$

$$BQ - HR = 0, \tag{5.4}$$

$$H'P = 0, \tag{5.5}$$

$$-H'Q = I_r. \tag{5.6}$$

These equations are easily solved since B is positive definite and H is of rank

r so that $H'B^{-1}H$ is non-singular. We obtain

$$R = -(H'B^{-1}H)^{-1},$$
$$Q = -BH(H'B^{-1}H)^{-1},$$
$$P = B^{-1}[I_s - H(H'B^{-1}H)^{-1}H'B^{-1}].$$

We note at this stage, though we do not require this result immediately, that the matrix P has rank $s - r$. For, from (5.5) since rank $(H') = r$, rank $(P) \leqq s - r$. While from (5.3) we have $s = \text{rank } (P - HQ') \leqq \text{rank } (P) + \text{rank } (HQ') \leqq \text{rank } (P) + r$, and so rank $(P) \geqq s - r$.

We return now to equations (5.1) and (5.2). If δ is sufficiently small then the matrix

$$\begin{bmatrix} B_{\theta_0} + \hat{b}(x) & -[H_{\theta_0} + \hat{h}(x)] \\ -[H'_{\theta_0} + \hat{h}^*(x)] & 0 \end{bmatrix}$$

also will be non-singular and we will write

$$\begin{bmatrix} B_{\theta_0} + \hat{b}(x) & -[H_{\theta_0} + \hat{h}(x)] \\ -[H'_{\theta_0} + \hat{h}^*(x)] & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \hat{P}(x) & \hat{Q}_1(x) \\ \hat{Q}_2(x) & \hat{R}(x) \end{bmatrix}.$$

Hence, from (5.1) and (5.2), for sufficiently small δ, if $x \in X_n$, we have

$$\begin{bmatrix} \hat{\theta}(x) - \theta_0 \\ \frac{1}{n}\hat{\lambda}(x) \end{bmatrix} = \begin{bmatrix} \hat{P}(x) & \hat{Q}_1(x) \\ \hat{Q}_2(x) & \hat{R}(x) \end{bmatrix} \begin{bmatrix} \frac{1}{n} l(x, \theta_0) \\ 0 \end{bmatrix}. \tag{5.7}$$

If the functions $\hat{\theta}$ and $\hat{\lambda}$ were defined for the whole of $\mathfrak{R}^n$ we could now discuss immediately the asymptotic distribution of these functions. However this is not necessarily so, and we go through the formality of extending the definition of these functions to the whole of $\mathfrak{R}^n$. We will then show that the random variables thus defined are asymptotically normally distributed and, in this sense, we may say that a consistent maximum likelihood estimator $\hat{\theta}$ of θ_0 is asymptotically normally distributed.

We let (δ_m), (ϵ_m) be decreasing sequences of positive numbers, such that $\epsilon_1 < 1$, $\delta_1 < \mu_0/\kappa_3$ (see Theorem 1), and $\delta_m \to 0$ and $\epsilon_m \to 0$ as $m \to \infty$. We then define an increasing sequence (n_m) of integers such that, if $n \geqq n_m$, there exists a set in $\mathfrak{R}^n$ with the properties $\mathfrak{A}1$ to $\mathfrak{A}4$ for $\epsilon = \epsilon_m$ and $\delta = \delta_m$. For $m = 1, 2, \cdots$, if $n_m \leqq n < n_{m+1}$ we choose a set X_n with the properties $\mathfrak{A}1$ to $\mathfrak{A}4$ for $\epsilon = \epsilon_m$ and $\delta = \delta_m$. Hence $\Pr\{X_n\} \to 1$ as $n \to \infty$ and if $n_m \leqq n < n_{m+1}$ and $x \in X_n$, the likelihood equations (2.1) and (2.2) have a solution $\hat{\theta}_n(x)$, $\hat{\lambda}_n(x)$ such that $\| \hat{\theta}_n(x) - \theta_0 \| < \delta_m$. Moreover for sufficiently large m, $\hat{\theta}_n(x)$ is a *maximum* likelihood estimate of θ_0, by §4. We now extend the definition of $\hat{\theta}_n$ and $\hat{\lambda}_n$ to $\mathfrak{R}^n$ by letting

$$\begin{bmatrix} \hat{\theta}_n(x) - \theta_0 \\ \frac{1}{n}\hat{\lambda}_n(x) \end{bmatrix} = \begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} \begin{bmatrix} \frac{1}{n} l(x, \theta_0) \\ 0 \end{bmatrix}, \quad \text{if} \quad x \in X_n.$$

We have thus defined sequences $(\hat{\theta}_n)$, $(\hat{\lambda}_n)$, $n = n_m, n_{m+1}, \cdots$ of random variables which have the property that θ_n converges in probability to θ_0 and with probability tending to 1 as $n \to \infty$, $\hat{\theta}_n$, $\hat{\lambda}_n$ satisfy the likelihood equations (2.1) and (2.2).

THEOREM 2. *The random variables* $n^{1/2}(\hat{\theta}_n - \theta_0)$, $n^{-1/2}\hat{\lambda}_n$ *are asymptotically jointly normally distributed with variance-covariance matrix*

$$\begin{bmatrix} P & 0 \\ 0 & -R \end{bmatrix}.$$

PROOF. If $x \in X_n$, we define $\hat{P}(x) = P$, $\hat{Q}_1(x) = Q$, $\hat{Q}_2(x) = Q'$ and $\hat{R}(x) = R$. Then for sufficiently large n, by (5.7) we may write

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \frac{1}{\sqrt{n}}\hat{\lambda}_n \end{bmatrix} = \begin{bmatrix} \hat{P} & \hat{Q}_1 \\ \hat{Q}_2 & \hat{R} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n}} l(\cdot, \theta_0) \\ 0 \end{bmatrix}.$$

The elements of the matrix

$$\begin{bmatrix} \hat{P} & \hat{Q}_1 \\ \hat{Q}_2 & \hat{R} \end{bmatrix}$$

are random variables which converge in probability to the corresponding elements of the matrix

$$\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix},$$

since in (5.1) and (5.2) $\hat{b}$, $\hat{h}$ and $\hat{h}^*$ tend to 0 as $\delta \to 0$. Also the s-dimensional random variable $n^{-1/2}l(\cdot, \theta_0)$ is asymptotically normally distributed with zero mean and variance-covariance matrix B_{θ_0} (Cramér [1]), and the $(s + r)$-dimensional random variable $(n^{-1/2}l(\cdot, \theta_0), 0)$ is asymptotically normally distributed with zero mean and variance-covariance matrix

$$\begin{bmatrix} B_{\theta_0} & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows by an extension, to a multi-dimensional random variable, of a theorem of Cramér [2], that $\sqrt{n}(\hat{\theta}_n - \theta_0)$, $n^{-1/2}\hat{\lambda}_n$ are jointly asymptotically normally distributed with zero mean and variance-covariance matrix.

$$\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} \begin{bmatrix} B_{\theta_0} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} = \begin{bmatrix} PB_{\theta_0}P & PB_{\theta_0}Q \\ Q'B_{\theta_0}P & Q'B_{\theta_0}Q \end{bmatrix}.$$

(We omit details of the proof of this extension though this result, in contrast to Cramér's result for real-valued random variables, is best obtained by considering characteristic functions). Now from (5.3), $PB_{\theta_0}P - PH_{\theta_0}Q' = P$. Since P is symmetric, $PH_{\theta_0} = P'H_{\theta_0} = 0$ by (5.5). Hence $PB_{\theta_0}P = P$. Similarly $PB_{\theta_0}Q = 0$ and $Q'B_{\theta_0}Q = -R$.

This completes the proof of the Theorem. We note, however, that, as might be expected, the asymptotic normal distribution of the s-dimensional random variable $n^{1/2}(\hat{\theta}_n - \theta_0)$ is improper, being by the note in Lemma 3 of rank $s - r$.

6. Numerical solution of likelihood equations. In this section we will discuss an iterative procedure for solving (2.1) and (2.2) numerically, which yields an estimate of the matrices P and R.

In any practical situation we do not know θ_0, and the only way in which we can verify that the conditions $\mathfrak{F}$ and $\mathfrak{C}$ are satisfied is to find that, for every θ belonging to some set U, *in which we know* θ_0 *lies*, the following conditions $\mathfrak{F}'$, $\mathfrak{C}'$ are satisfied.

$\mathfrak{F}'1$, $\mathfrak{F}'2$. For every $\theta \in U$, $\mathfrak{F}1$ and $\mathfrak{F}2$ are satisfied.

$\mathfrak{F}'3$ For every $\theta \in U$ and $i, j, k = 1, 2, \cdots, s$,

$$\left| \frac{\partial^3 \log f(t, \theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| < F_3(t)$$

and

$$\int_{-\infty}^{\infty} F_3(t) f(t, \theta)\, dt \leqq \kappa_1',$$

a finite number.

$\mathfrak{F}'4$. For every $\theta \in U$,

$$b_{ij} = \int_{-\infty}^{\infty} \frac{\partial \log f(t, \theta)}{\partial\theta_i} \frac{\partial \log f(t, \theta)}{\partial\theta_j} f(t, \theta)\, dt,$$

$i, j = 1, 2, \cdots, s$, are finite, the matrix $B_\theta = (b_{ij}(\theta))$ is positive definite and, if μ_θ is the minimum latent root of B_θ, then $\mu_\theta \geqq \mu_0'$ where μ_0' is a given number greater than 0.

$\mathfrak{C}'1$, $\mathfrak{C}'2$. For every $\theta \in U$, $\mathfrak{C}1$ and $\mathfrak{C}2$ are satisfied.

$\mathfrak{C}'3$ For every $\theta \in U$, H_θ is of rank r.

The conditions $\mathfrak{F}'$ are a straightforward generalization of Cramér's conditions [2].

We will now assume that the conditions $\mathfrak{F}'$ and $\mathfrak{C}'$ are satisfied, that x is such that the likelihood equations (2.1) and (2.2) have a solution $\hat{\theta}(x), \hat{\lambda}(x)$ and that $\theta^{(1)}$ is an initial approximation to $\hat{\theta}(x)$ such that $\| \theta^{(1)} - \hat{\theta}(x) \|$ is small. Then to a first order of approximation

$$l(x, \hat{\theta}) = l(x, \theta^{(1)}) + M_{x,\theta^{(1)}}(\hat{\theta} - \theta^{(1)}),$$

$$h(\hat{\theta}) = h(\theta^{(1)}) + H_{\theta^{(1)}}(\hat{\theta} - \theta^{(1)}).$$

Also if n is large, $(1/n)\hat{\lambda}(x)$ is near 0 for "most" x. We assume that x is a point for which $(1/n)\lambda(x)$ is near 0. Then we also have to a first order of approximation

$$H_{\hat{\theta}} \frac{1}{n} \hat{\lambda} = H_{\theta^{(1)}} \frac{1}{n} \hat{\lambda}.$$

Since $\hat{\theta}(x)$, $\hat{\lambda}(x)$ satisfy (2.1) and (2.2) then, approximately, we have

$$\begin{bmatrix} -\frac{1}{n} M_{x,\theta^{(1)}} & -H_{\theta^{(1)}} \\ -H'_{\theta^{(1)}} & 0 \end{bmatrix} \begin{bmatrix} \hat{\theta} - \theta^{(1)} \\ \frac{1}{n} \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} l(x, \theta^{(1)}) \\ h(\theta^{(1)}) \end{bmatrix}. \tag{6.1}$$

The normal situation, if n is large, is that $\hat{\theta}(x)$ is near θ_0. Consequently since $\theta^{(1)}$ is near $\hat{\theta}(x)$ the matrix $-(1/n)M_{x,\theta^{(1)}}$ approximates $-(1/n)M_{x,\theta_0}$ which in turn approximates B_{θ_0}. Then $B_{\theta^{(1)}}$ approximates B_{θ_0} and we propose to replace $-(1/n)M_{x,\theta^{(1)}}$ in (6.1) by $B_{\theta^{(1)}}$, and to obtain a correction to $\theta^{(1)}$, and an initial approximation to $(1/n)\lambda$, by solving the equation

$$\begin{bmatrix} B_{\theta^{(1)}} & -H_{\theta^{(1)}} \\ -H'_{\theta^{(1)}} & 0 \end{bmatrix} \begin{bmatrix} \hat{\theta} - \theta^{(1)} \\ \frac{1}{n} \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} l(x, \theta^{(1)}) \\ h(\theta^{(1)}) \end{bmatrix}. \tag{6.2}$$

The idea of replacing $-(1/n)M_{x,\theta^{(1)}}$ by $B_{\theta^{(1)}}$ is not original though the authors do not know where it originated.

Because of $\mathfrak{F}'4$, $\mathfrak{C}'3$, by Lemma 3, the matrix

$$\begin{bmatrix} B_{\theta^{(1)}} & -H_{\theta^{(1)}} \\ -H'_{\theta^{(1)}} & 0 \end{bmatrix}$$

is non-singular and we will denote its inverse by

$$\begin{bmatrix} P_1 & Q_1 \\ Q'_1 & R_1 \end{bmatrix}$$

We define $\theta^{(2)}$, $\lambda^{(2)}$ by

$$\begin{bmatrix} \theta^{(2)} \\ \frac{1}{n} \lambda^{(2)} \end{bmatrix} = \begin{bmatrix} \theta^{(1)} \\ 0 \end{bmatrix} + \begin{bmatrix} P_1 & Q_1 \\ Q'_1 & R_1 \end{bmatrix} \begin{bmatrix} \frac{1}{n} l(x, \theta^{(1)}) \\ h(\theta^{(1)}) \end{bmatrix}$$

and, more generally, $\theta^{(r)}$, $\lambda^{(r)}$ by (with the obvious definition of P_{r-1}, Q_{r-1} and R_{r-1}),

$$\begin{bmatrix} \theta^{(r)} \\ \frac{1}{n} \lambda^{(r)} \end{bmatrix} = \begin{bmatrix} \theta^{(r-1)} \\ 0 \end{bmatrix} + \begin{bmatrix} P_{r-1} & Q_{r-1} \\ Q'_{r-1} & R_{r-1} \end{bmatrix} \begin{bmatrix} \frac{1}{n} l(x, \theta^{(r-1)}) \\ h(\theta^{(r-1)}) \end{bmatrix}.$$

If the sequences $(\theta^{(r)})$, $(\lambda^{(r)})$ converge then they converge to a solution of the likelihood equations, as is easily verified. We do not attempt to give rigorous conditions under which these sequences do converge. However the fact that we may expect them to converge in most practical situations follows from the heuristic argument leading to (6.2).

We have thus established an iterative procedure for solving the likelihood equations. The heaviest part of the computation involved in this method is the inversion of a matrix and computation will normally be reduced by considering the sequences $(\hat{\theta}^{(r)})$, $(\hat{\lambda}^{(r)})$ defined by

$$\begin{bmatrix} \hat{\theta}^{(r)} \\ \frac{1}{n}\hat{\lambda}^{(r)} \end{bmatrix} = \begin{bmatrix} \hat{\theta}^{(r-1)} \\ \frac{1}{n}\hat{\lambda}^{(r-1)} \end{bmatrix} + \begin{bmatrix} P_1 & Q_1 \\ Q_1' & R_1 \end{bmatrix} \begin{bmatrix} \frac{1}{n} 1(x, \hat{\theta}^{(r-1)}) + H_{\hat{\theta}^{(r-1)}} \frac{1}{n}\lambda^{(r-1)} \\ h(\hat{\theta}^{(r-1)}) \end{bmatrix}$$

$r = 1, 2, \cdots$, where $\hat{\theta}^{(0)} = \theta^{(2)}$ and $\hat{\lambda}^{(0)} = \lambda^{(2)}$. Again if these sequences converge, they converge to a solution of the likelihood equations since

$$\begin{bmatrix} P_1 & Q_1 \\ Q_1' & R_1 \end{bmatrix}$$

is non-singular. And again we do not attempt to give conditions under which they do converge. The main justifications we put forward for this computational procedure are

(i) the similarity between this method and Newton's method, and

(ii) the fact that similar modifications of Newton's method have been used successfully elsewhere, for example in probit analysis [3]. The main advantage of this method of solving the likelihood equations is that it involves inversion of only one matrix.

7. Tests of the model. In a situation such as is outlined in §1 two natural questions arise in practice regarding the adequacy of the model introduced to describe an experimental situation.

(i) Does the true parameter point θ_0 satisfy the condition $h(\theta_0) = 0$?

(ii) Is the true parameter point some hypothetical point θ^* such that

$$h(\theta^*) = 0?$$

And this is the natural order for these questions since the second would be asked only if the first were answered in the affirmative. We now propose a procedure for answering these questions in this order.

(i) The most natural approach to the first question would be as follows. We would calculate an unrestrained maximum likelihood estimate $\hat{\theta}_u(x)$ of θ_0, and for $\hat{\theta}_u(x)$ we would have $\ell(x, \hat{\theta}_u(x)) = 0$. If $h(\hat{\theta}_u(x))$ were in some sense "near enough" $0 \in \mathfrak{R}^r$ then we would decide that in fact $h(\theta_0) = 0$. Dually, we might calculate a maximum likelihood estimate $\hat{\theta}(x)$ subject to the restraint

$$h(\hat{\theta}(x)) = 0$$

and then decide that $h(\theta_0) = 0$ if $\ell(x, \hat{\theta}(x))$ were "near enough" $0 \in \mathfrak{R}^s$. And the test we propose is based on the second possibility. We note that, by (2.1),

$$H_{\hat{\theta}}\hat{\lambda}(x) = -1(x, \hat{\theta}(x))$$

and it seems reasonable to decide that $h(\theta_0) = 0$ if $\hat{\lambda}(x)$ is in some sense 'near enough' $0 \in \mathfrak{R}^r$.

We have seen in Theorem 2 that when $h(\theta_0) = 0$, $n^{-1/2}\hat{\lambda}$ is normally distributed asymptotically with variance-covariance matrix $-R$, which is of rank r. Consequently $-(1/n)\hat{\lambda}'R^{-1}\hat{\lambda}$ is asymptotically distributed as χ^2 with r degrees of freedom, when $h(\theta_0) = 0$, and, in obvious notation, $-(1/n)\hat{\lambda}'R_{\hat{\theta}}^{-1}\hat{\lambda}$ also is approximately, for large n, distributed as χ^2 with r degrees of freedom. We propose to choose as a region of acceptance of the hypothesis that $h(\theta_0) = 0$ the set of x for which

$$-\frac{1}{n}\hat{\lambda}'(x)R_{\hat{\theta}(x)}^{-1}\hat{\lambda}(x) \leqq k,$$

where k is determined by

$$\Pr\{\chi_{[r]}^2 \leqq k\} = 0.95.$$

This gives a test of size 95% of the hypothesis that $h(\theta_0) = 0$.

(ii) The natural corollary of using the asymptotic distribution of $\hat{\lambda}$ in this way is to use the asymptotic distribution of $\hat{\theta}$ as established in Theorem 2 to answer the second question. If $\theta^* = \theta_0$ then $n(\hat{\theta} - \theta^*)'B_{\hat{\theta}^*}(\hat{\theta} - \theta^*)$ is approximately distributed as χ^2 with $s - r$ degrees of freedom if n is large. This is easily established by noting that a consequence of equations (5.3)–(5.6) is that $B^{-1} = PBP - QR^{-1}Q'$, and hence that

$$\frac{1}{n}l'B^{-1}l = n(\hat{\theta} - \theta_0)'B(\hat{\theta} - \theta_0) - \frac{1}{n}\hat{\lambda}'R^{-1}\hat{\lambda}.$$

We use this fact as in the previous paragraph to establish a region of acceptance of the hypothesis that the true parameter point is θ^*.

Here no attempt is made to justify this test on other than an intuitive basis. Since the Lagrangian multiplier test seems to be of wide applicability and of considerable importance in practical statistics, it will be fully discussed both from the theoretical and practical points of view in subsequent papers.

REFERENCES

[1] H. Cramér, "Random variables and probability distributions," *Cambridge University Press*, 1937.
[2] H. Cramér, "Mathematical methods of statistics," *Princeton University Press*, 1949.
[3] D. J. Finney, "Probit analysis," *Cambridge University Press*, 1947.
[4] S. Lefschetz, "Introduction to topology," *Princeton University Press*, 1949.

AITCHISON, J. and SILVEY, S.D. (1960)

Maximum-likelihood estimation procedures and associated tests of significance

Reprinted from *J. R. Statist. Soc.* B22, 154-71

Maximum-likelihood Estimation Procedures and Associated Tests of Significance

By J. AITCHISON and S. D. SILVEY
Department of Mathematics, University of Glasgow

[Received August, 1959]

SUMMARY

The essence of many statistical problems, including most standard techniques, is to test whether or not the unknown parameters of an appropriate statistical model satisfy certain restrictions; and the outcome of such a test dictates whether it is necessary to provide estimates of these parameters which also satisfy the restrictions. In this paper we discuss and illustrate the relative merits, as practical tools for the consulting statistician, of two large-sample techniques of wide applicability to such situations: (i) unrestricted maximum-likelihood estimation with its associated Wald test, (ii) restricted maximum-likelihood estimation with its associated Lagrange-multiplier test. The discussion falls into two main sections corresponding to two methods of specifying restrictions, as constraint equations in the parameters, or as freedom equations expressing the parameters in terms of a second smaller set of parameters. The methods are modified by a simple device to apply to the case where constraints on the parameters are necessary to allow their identification.

1. INTRODUCTION

WHEN translated into statistical terminology many of the problems put to statisticians by experimenters are of the following nature. Underlying distributions are known, except for a finite number of parameters, and the experimenter wishes to test a null hypothesis which states that the unknown parameters satisfy certain functional relationships. When this question has been answered he normally wishes estimates of some, or all, of the parameters and standard errors of these estimates. It is natural to demand that, when the null hypothesis is accepted, the estimates also satisfy the functional relationships; if, however, the null hypothesis is rejected, this demand would be absurd and estimates unrestricted by these relationships would then be appropriate.

Most of the problems which involve the use of standard techniques, such as the t-test, analysis of variance, the χ^2-test of homogeneity, fall into this category. Thus the t-test for the comparison of two means is used to answer the question whether two unknown parameters θ_1 and θ_2 satisfy the relation $\theta_1-\theta_2=0$. If the answer is in the affirmative, then an estimate of their common value is usually required. In a two-way classification analysis of variance the question "Is there no interaction?" is equivalent to the question "Do unknown means θ_{ij} $(i=1,2,\ldots,p;\ j=1,2,\ldots,q)$ satisfy $(p-1)(q-1)$ conditions of which a typical one is $\theta_{i_1j_1}-\theta_{i_1j_2}=\theta_{i_2j_1}-\theta_{i_2j_2}$?" Of course, the more usual way of posing the latter question is to express it in terms of freedom equations: can θ_{ij} be expressed in the form $\mu+\alpha_i+\beta_j$, where

$\alpha_1+\ldots+\alpha_p=0$ and $\beta_1+\ldots+\beta_q=0$? If so, estimates of μ, the α's and the β's are desired. Similarly, the question: is there homogeneity in a 2×2 contingency table? may be translated into: do two unknown probabilities θ_1 and θ_2 satisfy the condition $\theta_1-\theta_2=0$? There is no point in multiplying such familiar examples, though it is interesting to note that almost every standard situation that comes to mind yields a problem of the nature that we consider in this paper.

There are two large-sample techniques which, in the experience of the authors, allow the consulting statistician to face with more equanimity the approaches of an experimenter with an unfamiliar problem. The theory underlying these methods is given elsewhere, by Wald (1943), Aitchison and Silvey (1958) and Silvey (1959), but as so often happens, practical details of the methods are there obscured in a morass of mathematical conditions which are usually easily satisfied in practice. In this paper we will not be concerned with these conditions but will concentrate on the features of the methods which are of most interest to the practising statistician. From this point of view, the main outcome of the theory is that both methods yield tests which are usually large-sample equivalents of the likelihood-ratio test and estimators whose distributions are almost always asymptotically normal. Furthermore, standard tests are usually either exactly the tests given by one or other of these methods or small-sample refinements of these tests. While the techniques thus provide a unification of standard methods, it is in their application to non-standard situations that their primary interest lies.

As has already been indicated, the functional relationships among the parameters on which the null hypothesis is based may be specified in two ways: (i) in the form of constraint equations, (ii) in the form of freedom equations, or indeed even as a combination of constraint and freedom equations. The freedom equation specification involves new parameters which are of interest only if the null hypothesis is accepted. Which is the natural specification depends on the particular situation. Thus in the t-test discussed above it is natural to specify the single relation between θ_1 and θ_2 as the constraint equation $\theta_1-\theta_2=0$. But in the situation where we are asking whether the regression of y on x is linear, i.e. whether y-means $\theta_1, \theta_2, \ldots, \theta_p$, corresponding to different values $x_1, x_2, \ldots, x_p$ of x, lie on a straight line, it is much more natural to express the functional relations in terms of freedom equations and to put the question in the following form: can θ_i be expressed in the form $\theta_i=\alpha x_i+\beta$? The last question is of course equivalent to: do the θ's satisfy the $p-2$ restrictions $(\theta_{i+1}-\theta_i)/(x_{i+1}-x_i)=(\theta_{i+2}-\theta_{i+1})/(x_{i+2}-x_{i+1})$ $(i=1,2,\ldots,p-2)$? To any freedom equation specification there will correspond a constraint equation specification. However, it is not always easy to derive the latter from the former and, as the methods we are about to discuss are not equally suited to these different ways of specifying our functional relationships, we will deal separately with the different forms, discussing first the constraint equation specification.

2. Constraint Equation Specification of Restrictions and the Wald Method

The basic idea of this method, proposed by Wald (1943), is as follows. We calculate unrestricted maximum-likelihood estimates of the unknown parameters and then test the null hypothesis by asking if these estimates (which for large samples are likely to be near the corresponding true parameters) nearly enough satisfy the relationships which specify the null hypothesis. In this section we set out the practical details of the method.

Suppose we are given n independent observations $x_1, x_2, \ldots, x_n$ (which may be real or vector-valued), and the common underlying distribution of the x's is known except for s parameters $\theta_1, \theta_2, \ldots, \theta_s$. We denote by $\log L(x, \theta)$ the value of the log-likelihood function for given $x = (x_1, x_2, \ldots, x_n)$ at the point $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$. To find a maximum-likelihood‡ estimate $\theta^* = (\dot{\theta}_1, \dot{\theta}_2, \ldots, \dot{\theta}_s)$ of the true unknown parameter $\theta_0 = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_s^{(0)})$ we solve the unrestricted maximum-likelihood equations

$$\frac{\partial \log L(x, \theta)}{\partial \theta_i} = 0 \quad (i = 1, 2, \ldots, s). \tag{2.1}$$

Normally, in practice, these equations have to be solved numerically by an iterative process and generally a reasonably economical such process is as follows. We denote by $\mathbf{B}_\theta$ the information matrix, i.e. $\mathbf{B}_\theta$ is the $s \times s$ matrix whose (i,j)th element is

$$-\frac{1}{n} E_\theta \left(\frac{\partial^2 \log L(x, \theta)}{\partial \theta_i \partial \theta_j} \right),$$

where E_θ denotes expected value with respect to the distribution corresponding to the point θ. If $\theta^{(1)}$ is an initial approximation to the solution of equations (2.1), then usually a better approximation $\theta^{(2)}$ is given by

$$\theta^{(2)} = \theta^{(1)} + \frac{1}{n} \mathbf{B}_{\theta^{(1)}}^{-1} \mathbf{D} \log L(x, \theta^{(1)}),$$

in matrix notation. Here $\mathbf{D} \log L$ denotes the s-dimensional column vector whose ith component is $\partial \log L / \partial \theta_i$.

Successive approximations to θ^* are $\theta^{(3)}, \theta^{(4)}, \ldots$, where

$$\theta^{(k+1)} = \theta^{(k)} + \frac{1}{n} \mathbf{B}_{\theta^{(1)}}^{-1} \mathbf{D} \log L(x, \theta^{(k)}).$$

The heaviest part of the computation involved in solving the equations (2.1) is the inversion of the matrix $\mathbf{B}_{\theta^{(1)}}$. We emphasize, however, that the inversion of only *one* matrix is required because $\mathbf{B}_{\theta^{(1)}}^{-1}$ is used in all iterations. Also the matrix $\mathbf{B}_{\theta^{(1)}}^{-1}/n$ may be used as an estimate of the variance matrix of $\dot{\theta}_1, \dot{\theta}_2, \ldots, \dot{\theta}_s$, though $\mathbf{B}_{\theta^*}^{-1}/n$ is a better such estimate.

All this is standard maximum-likelihood theory which we have reproduced for the sake of completeness and in order to establish the notation of the paper.

We now suppose that our null hypothesis $\mathscr{H}_0$ states that the unknown parameters satisfy $r (< s)$ well-behaved relationships which we write in the form

$$h_1(\theta) = h_2(\theta) = \ldots = h_r(\theta) = 0.$$

We denote by $\mathbf{h}(\theta)$ the r-dimensional column vector whose ith component is $h_i(\theta)$ and by $\mathbf{H}_\theta$ the $s \times r$ matrix whose (i,j)th element is $\partial h_j(\theta)/\partial \theta_i$. If the restrictions are sensible ones which involve no redundancy, $\mathbf{H}_\theta$ will be of rank r and we assume this to be the case. Wald's method of testing $\mathscr{H}_0$ uses the statistic

$$n \mathbf{h}'(\theta^*) [\mathbf{H}'_{\theta^*} \mathbf{B}_{\theta^*}^{-1} \mathbf{H}_{\theta^*}]^{-1} \mathbf{h}(\theta^*).$$

Under the null hypothesis this is distributed as $\chi^2_{[r]}$ and we accept or reject $\mathscr{H}_0$ according as the value obtained is less than or greater than an appropriate upper percentage point of such a χ^2-distribution.

‡ For typographical reasons it has been necessary to use an asterisk in a sense different from that of previous papers on this subject.

3. Constraint Equation Specification of Restrictions and the Lagrange-multiplier Method

This method of testing the null hypothesis is based on the idea of asking if restricted estimates of the unknown parameters nearly enough give an absolute maximum of the likelihood function. The initial part of the computation consists of finding estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$ which maximize $\log L(x, \theta)$ subject to the conditions $h_1(\theta) = h_2(\theta) = \ldots = h_r(\theta) = 0$. Normally this involves solving the restricted likelihood equations

$$\frac{1}{n}\frac{\partial \log L(x,\theta)}{\partial \theta_i} + \sum_{j=1}^{r} \lambda_j \frac{\partial h_j(\theta)}{\partial \theta_i} = 0 \quad (i = 1, 2, \ldots, s)$$

$$h_j(\theta) = 0 \quad (j = 1, 2, \ldots, r), \tag{3.1}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_r$ are Lagrange multipliers.

Again, in practice these equations generally have to be solved numerically and there is an iterative method of solving them similar to that used for solving the equations (2.1).

Suppose $\theta^{(1)}$ is an initial approximation to $\theta\dagger = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s)$. Then if

$$\begin{bmatrix} \theta^{(2)} \\ \lambda^{(2)} \end{bmatrix} = \begin{bmatrix} \theta^{(1)} \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_{\theta^{(1)}} & -\mathbf{H}_{\theta^{(1)}} \\ -\mathbf{H}'_{\theta^{(1)}} & 0 \end{bmatrix}^{-1} \begin{bmatrix} (1/n)\mathbf{D}\log L(x, \theta^{(1)}) \\ \mathbf{h}(\theta^{(1)}) \end{bmatrix},$$

it is usually the case that $\theta^{(2)}$ and $\lambda^{(2)}$ give a better approximation than $\theta^{(1)}$ and 0 to the solution of the equations. Successive approximations $(\theta^{(3)}, \lambda^{(3)}), (\theta^{(4)}, \lambda^{(4)}), \ldots$ are given by

$$\begin{bmatrix} \theta^{(k+1)} \\ \lambda^{(k+1)} \end{bmatrix} = \begin{bmatrix} \theta^{(k)} \\ \lambda^{(k)} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_{\theta^{(1)}} & -\mathbf{H}_{\theta^{(1)}} \\ -\mathbf{H}'_{\theta^{(1)}} & 0 \end{bmatrix}^{-1} \begin{bmatrix} (1/n)\mathbf{D}\log L(x, \theta^{(k)}) + \mathbf{H}_{\theta^{(k)}}\lambda^{(k)} \\ \mathbf{h}(\theta^{(k)}) \end{bmatrix}. \tag{3.2}$$

It is worth remarking here that if any restriction $h_j(\theta) = 0$ is linear in θ, then $h_j(\theta^{(k)}) = 0$ implies that $h_j(\theta^{(k+1)}) = 0$, i.e. if at any stage a linear restriction is satisfied, it is satisfied at every subsequent stage.

As before, the main computational task in solving equations (3.1) is the inversion of a matrix, this time of order $(s+r)\times(s+r)$. Only one such inversion is involved, however, and again the process of solving the equations yields as a by-product an estimate of the variance matrix of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$. For if

$$\begin{bmatrix} \mathbf{B}_\theta & -\mathbf{H}_\theta \\ -\mathbf{H}'_\theta & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P}_\theta & \mathbf{Q}_\theta \\ \mathbf{Q}'_\theta & \mathbf{R}_\theta \end{bmatrix},$$

then $\mathbf{P}_{\theta^{(1)}}/n$ estimates the variance matrix of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$, though $\mathbf{P}_{\theta\dagger}/n$ is a better such estimate.

The Lagrange-multiplier method of testing the null hypothesis is based on the statistic

$$-n\hat{\lambda}\mathbf{R}^{-1}_{\theta\dagger}\hat{\lambda} = n(\mathbf{H}_{\theta\dagger}\hat{\lambda})'\mathbf{B}^{-1}_{\theta\dagger}(\mathbf{H}_{\theta\dagger}\hat{\lambda}) = \frac{1}{n}[\mathbf{D}\log L(x, \theta\dagger)]'\mathbf{B}^{-1}_{\theta\dagger}[\mathbf{D}\log L(x, \theta\dagger)].$$

For large samples this is distributed, under $\mathcal{H}_0$, as $\chi^2_{[r]}$, and we accept or reject $\mathcal{H}_0$ according as the value of this statistic is less than or greater than an appropriate upper percentage point of a $\chi^2_{[r]}$ distribution.

4. A Useful Method of Matrix Inversion

Seldom in practice is there difficulty in inverting $\mathbf{B}_{\theta^{(1)}}$ because of its low order or its simplicity, e.g. it may be diagonal. We therefore concentrate on describing a computational routine for the inversion involved in the Lagrange-multiplier method.

Computational routine. To compute

$$\begin{bmatrix} \mathbf{B} & -\mathbf{H} \\ -\mathbf{H}' & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix}$$

we first obtain $\mathbf{B}^{-1}$ and then carry out the following steps:

1. Compute $\mathbf{H}'\mathbf{B}^{-1}$ and $\mathbf{H}'\mathbf{B}^{-1}\mathbf{H}$.
2. Invert $\mathbf{H}'\mathbf{B}^{-1}\mathbf{H}$, obtaining a matrix $-\mathbf{R}$.
3. Compute $\mathbf{Q}' = \mathbf{R}(\mathbf{H}'\mathbf{B}^{-1})$.
4. Compute $\mathbf{P} = \mathbf{B}^{-1} + \mathbf{Q}(\mathbf{H}'\mathbf{B}^{-1})$; the symmetry of this last matrix provides a useful partial check on the computation.

Apart from step 2 the method involves only the relatively easy operations of matrix multiplication and addition. The matrix $\mathbf{H}'\mathbf{B}^{-1}\mathbf{H}$ is of order $r \times r$ and is easily inverted in many applications since the number r of restrictions is often small.

When we feel it is necessary to improve our first estimate of the variance matrix we may, as an alternative to the above method of matrix inversion, use an iterative scheme, starting with our initial inverse based on $\theta^{(1)}$, to obtain the final inverse based on $\theta\dagger$. One such scheme uses $\mathbf{T}_1$ as an approximation to $\mathbf{S}^{-1}$ to generate successive approximations $\mathbf{T}_2, \mathbf{T}_3, \ldots$ to $\mathbf{S}^{-1}$ by $\mathbf{T}_{k+1} = \mathbf{T}_k(2\mathbf{I} - \mathbf{S}\mathbf{T}_k)$, a process which requires only matrix multiplication.

5. Assessment of the Two Methods

Before attempting to compare the methods we provide a simple example to illustrate what the two tests yield in a standard problem.

Example 1: test of a proportion. In a given population the proportion θ of individuals possessing a certain attribute is unknown and we wish to test the hypothesis that $\theta = p$, so that $\mathscr{H}_0$ is specified by $h(\theta) = \theta - p = 0$. In a large random sample (with replacement) of n individuals m are found to possess the attribute. Here the common distribution of the x's involves one unknown parameter θ and is specified by $\mathrm{pr}\{x = 1 \mid \theta\} = \theta = 1 - \mathrm{pr}\{x = 0 \mid \theta\}$. Also $m = x_1 + \ldots + x_n$. Then

$$\log L(x, \theta) = \text{const} + m \log \theta + (n - m) \log (1 - \theta),$$

$$\mathbf{D} \log L(x, \theta) = \frac{m - n\theta}{\theta(1 - \theta)}, \quad \mathbf{B}_\theta = \frac{1}{\theta(1 - \theta)}, \quad \mathbf{H}_\theta = 1.$$

We also clearly have $\hat{\theta} = m/n$ and $\hat{\theta} = p$.

The Wald test statistic is then

$$\frac{n(\hat{\theta} - p)^2}{\hat{\theta}(1 - \hat{\theta})} = \frac{(m - np)^2}{n(m/n)(1 - m/n)}$$

and the Lagrange-multiplier test statistic

$$\frac{1}{n} \frac{(m - n\hat{\theta})^2}{\hat{\theta}(1 - \hat{\theta})} = \frac{(m - np)^2}{np(1 - p)},$$

and both are to be compared with $\chi^2_{[1]}$ points. The latter is a familiar large-sample binomial test and the former is an obvious large-sample equivalent.

The question of which of the two general methods is more efficient for dealing with any practical problem is a nicely balanced one, and it is impossible to give any hard and fast rule in this connection. The computation involved in applying Wald's test automatically yields unrestricted estimates of the unknown parameters and their variance matrix; and this is what we want if the null hypothesis is rejected. On the other hand, the Lagrange-multiplier test yields restricted estimates and their variances, which we want if $\mathscr{H}_0$ is accepted. If we knew *a priori* whether or not $\mathscr{H}_0$ were going to be accepted, then we could obviously choose between the methods. This we never know in any given situation, though it may be possible to make a good guess based on a careful scrutiny of the data. Moreover, particular situations may have aspects which make one approach easier than the other. For instance, it is possible, and in fact happens quite regularly, that the unrestricted likelihood equations can be solved directly to obtain explicit formulae for the unrestricted estimates, whereas numerical solution of the restricted equations is necessary. In this case Wald's test of $\mathscr{H}_0$ is relatively easy to apply and it seems economical to use this test and then, if $\mathscr{H}_0$ is accepted, to solve the restricted equations. Indeed, the work employed in carrying out Wald's test can form the basis of the matrix inversion involved in solving these restricted equations. For we would have already computed $\mathbf{H}'_{\theta^*}\mathbf{B}^{-1}_{\theta^*}$ and $(\mathbf{H}'_{\theta^*}\mathbf{B}^{-1}_{\theta^*}\mathbf{H}_{\theta^*})^{-1}$ and so, if we take $\theta^{(1)} = \theta^*$ as our initial approximation to $\theta\dagger$, we would have already carried out the first three steps and hence the most difficult part of the suggested matrix inversion routine. Another slight advantage is that $\mathrm{D}\log L(x, \theta^{(1)}) = 0$ and so $\theta^{(2)}, \lambda^{(2)}$ can be obtained as soon as $\mathbf{R}_{\theta^{(1)}}$ and $\mathbf{Q}_{\theta^{(1)}}$ have been found.

Thus often in practice a judicious combination of the methods involves the minimum of calculation. It is of some interest to compare the relative efficiencies of these methods and the direct application of the generalized likelihood-ratio test, which uses the statistic $\Lambda(x) = L(x, \theta^*)/L(x, \theta\dagger)$. For large samples, in the type of situation we have been discussing, $2\log\Lambda$ is distributed, under $\mathscr{H}_0$, as $\chi^2_{[r]}$ approximately. Thus to apply the likelihood-ratio test we have to calculate both θ^* and $\theta\dagger$, whereas either the Wald test or the Lagrange-multiplier test involves the calculation of only one of these estimates. It is true that either of the foregoing techniques for testing $\mathscr{H}_0$ with subsequent estimation *may* involve calculation of both θ^* and $\theta\dagger$. But since under either of these techniques we *sometimes* have to calculate only one estimate, and since the calculation of estimates usually forms the heaviest part of the total computation, both the present techniques seem preferable to straightforward application of the likelihood-ratio test.

To illustrate these points we will consider two applications. The first is straightforward and the second involves a mild extension of the techniques designed to meet a difficulty encountered in practice.

Example 2. The mean of 50 independent observations from a normal distribution is 4·6 and their standard deviation is 4·0. A null hypothesis $\mathscr{H}_0$ states that the standard deviation of the distribution is equal to its mean. If $\mathscr{H}_0$ is accepted, an estimate of the common value of the mean and standard deviation is desired, together with the standard error of this estimate.

This is a very simple problem and we introduce it in order to illustrate the techniques without extraneous difficulties.

There are two unknown parameters, the mean θ_1 and the standard deviation θ_2 of the underlying normal distribution. The hypothesis $\mathscr{H}_0$ says that the true values $\theta_1^{(0)}$ and $\theta_2^{(0)}$ of these parameters satisfy the single restriction $\theta_1 - \theta_2 = 0$. Now

$$\log L(x, \theta) = \text{const} - n \log \theta_2 - \frac{n}{2\theta_2^2}[s^2 + (\bar{x} - \theta_1)^2],$$

where n is the number of observations, $\bar{x}$ is the mean of the observations x, and s^2 is the variance of the observations, $\Sigma(x_i - \bar{x})^2/n$; and it is not difficult to verify that

$$\text{(i)} \quad \frac{1}{n}\mathbf{D}\log L(x, \theta) = \begin{bmatrix} (\bar{x} - \theta_1)/\theta_2^2 \\ -\frac{1}{\theta_2} + \frac{1}{\theta_2^3}\{s^2 + (\bar{x} - \theta_1)^2\} \end{bmatrix},$$

$$\text{(ii)} \quad \mathbf{B}_\theta^{-1} = \theta_2^2 \begin{bmatrix} 1 & 0 \\ 0 & 0{\cdot}5 \end{bmatrix}, \qquad \text{(iii)} \quad \mathbf{H}_\theta = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Here we can easily apply Wald's test immediately, for we can write down $\dot{\theta}_1 = \bar{x}$ and $\dot{\theta}_2 = s$. We then have $\mathbf{h}(\theta^*) = \bar{x} - s$ (a scalar because there is just one restriction), and $\mathbf{H}'_{\theta^*}\mathbf{B}_{\theta^*}^{-1}\mathbf{H}_{\theta^*} = \frac{3}{2}s^2$. From this it follows that

$$n\mathbf{h}'(\theta^*)[\mathbf{H}'_{\theta^*}\mathbf{B}_{\theta^*}^{-1}\mathbf{H}_{\theta^*}]^{-1}\mathbf{h}(\theta^*) = \frac{n(\bar{x} - s)^2}{\frac{3}{2}s^2} = 0{\cdot}75$$

for the given values of $\bar{x}$ and s. Since this is smaller than the upper 5 per cent value of a $\chi^2_{(1)}$ distribution, we accept $\mathscr{H}_0$.

To obtain an estimate of the common value of $\theta_1^{(0)}$ and $\theta_2^{(0)}$ we have to solve the restricted equations

$$\begin{aligned} \frac{\bar{x} - \theta_1}{\theta_2^2} + \lambda_1 &= 0, \\ -\frac{1}{\theta_2} + \frac{1}{\theta_2^3}[s^2 + (\bar{x} - \theta_1)^2] - \lambda_1 &= 0, \\ \theta_1 - \theta_2 &= 0. \end{aligned} \qquad (5.1)$$

and

In this very simple situation one would usually eliminate λ_1 and θ_2, and obtain the quadratic equation

$$\theta_1^2 + \bar{x}\theta_1 - (\bar{x}^2 + s^2) = 0,$$

of which $\hat{\theta}_1 = \hat{\theta}_2$ is the positive root, 4·21 for $\bar{x} = 4{\cdot}6$ and $s = 4{\cdot}0$. However, for the sake of illustration we will carry through the iterative process of solving these equations outlined in section 3. As indicated earlier, we can simply take as initial approximations to the restricted estimates the unrestricted estimates already found. Hence, with $\theta_1^{(1)} = 4{\cdot}6$ and $\theta_2^{(1)} = 4{\cdot}0$, we have

$$\begin{bmatrix} \mathbf{B}_{\theta^{(1)}} & -\mathbf{H}_{\theta^{(1)}} \\ -\mathbf{H}'_{\theta^{(1)}} & \mathbf{0} \end{bmatrix}^{-1} = \frac{1}{24}\begin{bmatrix} 128 & 128 & -16 \\ 128 & 128 & 8 \\ -16 & 8 & -1 \end{bmatrix},$$

$$\begin{bmatrix} \frac{1}{n}\mathbf{D}\log L(x, \theta^{(1)}) \\ \mathbf{h}(\theta^{(1)}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0{\cdot}6 \end{bmatrix}.$$

Hence second approximations $\theta_1^{(2)}$, $\theta_2^{(2)}$ and $\lambda_1^{(2)}$ are given by

$$\begin{bmatrix} \theta_1^{(2)} \\ \theta_2^{(2)} \\ \lambda_1^{(2)} \end{bmatrix} = \begin{bmatrix} 4{\cdot}6 \\ 4{\cdot}0 \\ 0 \end{bmatrix} + \frac{1}{24} \begin{bmatrix} 128 & 128 & -16 \\ 128 & 128 & 8 \\ -16 & 8 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0{\cdot}6 \end{bmatrix} = \begin{bmatrix} 4{\cdot}2 \\ 4{\cdot}2 \\ -0{\cdot}025 \end{bmatrix}.$$

Then

$$\begin{bmatrix} \frac{1}{n} \mathbf{D} \log L(x, \theta^{(2)}) + \mathbf{H}_{\theta^{(2)}} \lambda^{(2)} \\ \mathbf{h}(\theta^{(2)}) \end{bmatrix} = \begin{bmatrix} -0{\cdot}0023 \\ 0{\cdot}0050 \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} \theta_1^{(3)} \\ \theta_2^{(3)} \\ \lambda_1^{(3)} \end{bmatrix} = \begin{bmatrix} 4{\cdot}2 \\ 4{\cdot}2 \\ -0{\cdot}025 \end{bmatrix} + \frac{1}{24} \begin{bmatrix} 128 & 128 & -16 \\ 128 & 128 & 8 \\ -16 & 8 & -1 \end{bmatrix} \begin{bmatrix} -0{\cdot}0023 \\ 0{\cdot}0050 \\ 0 \end{bmatrix} = \begin{bmatrix} 4{\cdot}21 \\ 4{\cdot}21 \\ -0{\cdot}022 \end{bmatrix},$$

which is a sufficiently close approximation to the exact solution of the equations (5.1).

We note, in passing, that the value of the statistic on which the Lagrange-multiplier test is based, viz. $n(\mathbf{H}_{\theta\dagger}\hat{\lambda})' \mathbf{B}_{\theta\dagger}^{-1}(\mathbf{H}_{\theta\dagger}\hat{\lambda})$, is, in this case, 0·64, a value which, as we might expect, results in the acceptance of $\mathscr{H}_0$; and, for the likelihood-ratio test the value of $2 \log \Lambda$ is 0·69, again resulting in the acceptance of $\mathscr{H}_0$.

Finally, the variance matrix of $\hat{\theta}_1, \hat{\theta}_2$ is estimated by the leading 2×2 submatrix of

$$\frac{1}{50} \begin{bmatrix} \frac{1}{4{\cdot}21^2} & 0 & -1 \\ 0 & \frac{2}{4{\cdot}21^2} & 1 \\ -1 & 1 & 0 \end{bmatrix}^{-1},$$

which is

$$\frac{1}{50} \frac{4{\cdot}21^2}{3} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

If we compare this with an estimate of the variance matrix of $\dot{\theta}_1, \dot{\theta}_2$, which is

$$\frac{1}{n} \mathbf{B}_{\theta^*}^{-1} = \frac{1}{50} 4^2 \begin{bmatrix} 1 & 0 \\ 0 & 0{\cdot}5 \end{bmatrix},$$

we obtain an idea of how use of the knowledge that the unknown parameter satisfies the restrictions specifying the null hypothesis reduces the variance of estimates. This is always so. We can use the property that restricted estimators are asymptotically normally distributed to obtain in the usual way an approximate confidence interval for $\theta_1^{(0)} = \theta_2^{(0)}$; with confidence coefficient 95 per cent. this is (3·54, 4·88).

6. Singular Information Matrices

Both of the foregoing techniques depend essentially on the non-singularity of the information matrix and this non-singularity is related to identifiability of the unknown parameters. Generally, in a situation where the unknown parameters are identifiable,

the information matrix is non-singular, and so it would seem that in practice singularity of the information matrix will not arise. However, this is not so, because it is sometimes convenient in describing an experiment to introduce unidentifiable parameters and then to impose restrictions on these parameters so that the restricted parameters are identifiable. Thus in a two-way classification analysis of variance situation with no interaction it is usual to express the mean of the (i,j)th class in the form $\mu+\alpha_i+\beta_j$, and these parameters are not identifiable unless some restrictions such as $\Sigma\alpha_i = 0$ and $\Sigma\beta_j = 0$ are imposed. Similarly, when the underlying distribution is multinomial (a fertile source of problems of the nature we are discussing) it is natural, for reasons of symmetry, to use the following description. Suppose there are s classes in which each observation may fall. Then we introduce s parameters $\theta_1, \theta_2, \ldots, \theta_s$ and denote by $\theta_i/(\theta_1+\theta_2+\ldots+\theta_s)$ the probability that a result will fall in the ith class. With this description the s parameters are not identifiable and the matrix $\mathbf{B}_\theta$ is singular. If we impose the restriction $\theta_1+\ldots+\theta_s = 1$, the resulting restricted parameters are identifiable, but $\mathbf{B}_\theta$ remains singular and the techniques so far developed break down.

In any particular case we can obviously overcome this difficulty by using the restrictions necessary for identifiability to eliminate certain of the parameters from the model, leaving a smaller number of parameters which are identifiable and an information matrix of correspondingly smaller order which is non-singular. Thus in the multinomial case we might say that the probabilities associated with the s classes are $\theta_1, \theta_2, \ldots, \theta_{s-1}$ and $1-\theta_1-\theta_2-\ldots-\theta_{s-1}$. We then have $s-1$ parameters which are identifiable and the information matrix of these parameters is non-singular. But this solution is unsatisfactory if only because of its lack of symmetry; and there is a much tidier method of overcoming the difficulty which we now discuss.

In general, we suppose that we have s unknown parameters $\theta_1, \theta_2, \ldots, \theta_s$ and a total of r restrictions $h_1(\theta) = h_2(\theta) = \ldots h_r(\theta) = 0$. Now, however, exactly $t(<r)$ of these restrictions are necessary to make all s parameters identifiable and the remaining $r-t$ restrictions define the null hypothesis. Usually the fact that t restrictions are necessary for identifiability corresponds to the fact that the information matrix is of rank $s-t$. Without any loss of generality we may suppose that the *first* t restrictions are necessary and sufficient for identifiability. This is normally reflected in the fact that if the matrix $\mathbf{H}_\theta$ (i.e. the $s\times r$ matrix whose (i,j)th element is $\partial h_j(\theta)/\partial\theta_i$) is partitioned into $[\mathbf{H}_{1\theta}\,\mathbf{H}_{2\theta}]$, where $\mathbf{H}_{1\theta}$ is of order $s\times t$, then $\mathbf{B}_\theta+\mathbf{H}_{1\theta}\mathbf{H}'_{1\theta}$ will be non-singular. This is the key to the method of adapting our techniques. Roughly speaking, all we need do is to replace the matrix $\mathbf{B}_\theta$ wherever it appears in the preceding theory by $\mathbf{B}_\theta+\mathbf{H}_{1\theta}\mathbf{H}'_{1\theta}$.

Looking at this in more detail, we have to alter our interpretation of unrestricted maximum-likelihood estimates $\hat\theta_1, \hat\theta_2, \ldots, \hat\theta_s$. These are now estimates which maximize the likelihood function subject to the identifiability conditions

$$h_1(\theta) = h_2(\theta) = \ldots = h_t(\theta) = 0,$$

and emerge as a solution of the equations

$$\frac{1}{n}\frac{\partial \log L(x,\theta)}{\partial\theta_i} + \sum_{j=1}^{t}\nu_j\frac{\partial h_j(\theta)}{\partial\theta_i} = 0 \quad (i = 1,2,\ldots,s)$$

$$h_j(\theta) = 0 \quad (j = 1,2,\ldots,t), \qquad (6.1)$$

where $\nu_1, \nu_2, \ldots, \nu_t$ are Lagrange multipliers.

Restricted maximum-likelihood estimates will, as before, usually emerge as solutions of the equations

$$\frac{1}{n}\frac{\partial \log L(x,\theta)}{\partial \theta_i}+\sum_{j=1}^{r}\lambda_j\frac{\partial h_j(\theta)}{\partial \theta_i}=0 \quad (i=1,2,\ldots,s),$$

$$h_j(\theta)=0 \quad (j=1,2,\ldots,r). \tag{6.2}$$

Both sets, (6.1) and (6.2), of equations, are of the type considered in section 3. The method of solution there considered depends on the positive definiteness of $\mathbf{B}_\theta$ and so it is not directly applicable in this new situation. But the device of replacing $\mathbf{B}_\theta$ by $\mathbf{B}_\theta+\mathbf{H}_{1\theta}\mathbf{H}'_{1\theta}$ is the only adaptation required to overcome this difficulty. Thus successive approximations $(\theta^{(k)},\nu^{(k)})$, $k=2,3,\ldots$, to the solution $(\theta^*,\dot{\nu})$ of (6.1) are given by

$$\begin{bmatrix}\theta^{(k+1)}\\ \nu^{(k+1)}\end{bmatrix}=\begin{bmatrix}\theta^{(k)}\\ \nu^{(k)}\end{bmatrix}+\begin{bmatrix}\mathbf{B}_{\theta^{(k)}}+\mathbf{H}_{1\theta^{(k)}}\mathbf{H}'_{1\theta^{(k)}} & -\mathbf{H}_{1\theta^{(k)}}\\ -\mathbf{H}'_{1\theta^{(k)}} & \mathbf{0}\end{bmatrix}^{-1}\begin{bmatrix}\frac{1}{n}\mathbf{D}\log L(x,\theta^{(k)})+\mathbf{H}_{1\theta^{(k)}}\nu^{(k)}\\ \mathbf{h}_1(\theta^{(k)})\end{bmatrix}$$

where $\mathbf{h}_1(\theta)$ is the t-dimensional column vector with ith component $h_i(\theta)$ $(i=1,2,\ldots,t)$. Similarly, the equations connecting successive approximations to the solution $(\theta\dagger,\hat{\boldsymbol{\lambda}})$ of (6.2) are simply the equations (3.2) with $\mathbf{B}_{\theta^{(i)}}+\mathbf{H}_{1\theta^{(i)}}\mathbf{H}'_{1\theta^{(i)}}$ replacing $\mathbf{B}_{\theta^{(i)}}$.

The modifications required in the tests of $\mathscr{H}_0$ are no more complicated. We recall that $\mathscr{H}_0$ now says: the unknown parameter θ_0 satisfies the $r-t$ restrictions $h_{t+1}(\theta)=h_{t+2}(\theta)=\ldots=h_r(\theta)=0$. The statistics used for testing $\mathscr{H}_0$ are obtained from those previously used by replacing $\mathbf{B}_\theta$ by $\mathbf{B}_\theta+\mathbf{H}_{1\theta}\mathbf{H}'_{1\theta}$ and now they are distributed, under $\mathscr{H}_0$, as $\chi^2_{[r-t]}$. Thus the Wald test statistic becomes

$$n\mathbf{h}'(\theta^*)\,[\mathbf{H}'_{\theta^*}(\mathbf{B}_{\theta^*}+\mathbf{H}_{1\theta^*}\mathbf{H}'_{1\theta^*})^{-1}\mathbf{H}_{\theta^*}]^{-1}\mathbf{h}(\theta^*)$$

and the Lagrange-multiplier statistic is changed to

$$n(\mathbf{H}_{\theta\dagger}\hat{\boldsymbol{\lambda}})'(\mathbf{B}_{\theta\dagger}+\mathbf{H}_{1\theta\dagger}\mathbf{H}'_{1\theta\dagger})^{-1}(\mathbf{H}_{\theta\dagger}\hat{\boldsymbol{\lambda}})$$

$$=\frac{1}{n}[\mathbf{D}\log L(x,\theta\dagger)]'\,[\mathbf{B}_{\theta\dagger}+\mathbf{H}_{1\theta\dagger}\mathbf{H}'_{1\theta\dagger}]^{-1}\,[\mathbf{D}\log L(x,\theta\dagger)].$$

Finally, estimates of the variance matrices of $\dot{\theta}_1,\dot{\theta}_2,\ldots,\dot{\theta}_s$ and $\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_s$ are now obtained as follows. Suppose

$$\begin{bmatrix}\mathbf{B}_\theta+\mathbf{H}_{1\theta}\mathbf{H}'_{1\theta} & -\mathbf{H}_{1\theta}\\ -\mathbf{H}'_{1\theta} & \mathbf{0}\end{bmatrix}^{-1}=\begin{bmatrix}\mathbf{U}_\theta & \mathbf{V}_\theta\\ \mathbf{V}_\theta & \mathbf{W}_\theta\end{bmatrix}$$

and

$$\begin{bmatrix}\mathbf{B}_\theta+\mathbf{H}_{1\theta}\mathbf{H}'_{1\theta} & -\mathbf{H}_\theta\\ -\mathbf{H}'_\theta & \mathbf{0}\end{bmatrix}^{-1}=\begin{bmatrix}\mathbf{P}^1_\theta & \mathbf{Q}^1_\theta\\ \mathbf{Q}^{1\prime}_\theta & \mathbf{R}^1_\theta\end{bmatrix}.$$

Then $\mathbf{U}_{\theta^*}/n$ is an estimate of the variance matrix of $\dot{\theta}_1,\dot{\theta}_2,\ldots,\dot{\theta}_s$, and $\mathbf{P}^1_{\theta\dagger}/n$ is an estimate of the variance matrix of $\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_s$. As before, approximations to these variance matrices arise automatically in the iterative procedure suggested for solving the restricted likelihood equations.

Example 3. Each individual in a population possesses one of the possible combinations of three attributes which we denote by a, b and c; i.e. each individual is

in one of seven classes characterized as a, b, c, ab, bc, ca and abc. We are interested in testing the null hypothesis that the attributes are independent in the sense that, for example, the probability that an individual chosen at random will fall in the class ab is the product of the probability that he will fall in a and the probability that he will fall in b. If we represent the probabilities associated with the classes (in the above order) by $\theta_i / \sum_{i=1}^{7} \theta_i$ $(i = 1, 2, ..., 7)$, then we are interested in the following restrictions:

$$h_1(\theta) = \sum_{i=1}^{7} \theta_i - 1 = 0, \qquad h_2(\theta) = \theta_4 - \theta_1 \theta_2 = 0,$$

$$h_3(\theta) = \theta_5 - \theta_2 \theta_3 = 0, \qquad h_4(\theta) = \theta_6 - \theta_3 \theta_1 = 0,$$

and

$$h_5(\theta) = \theta_7 - \theta_1 \theta_2 \theta_3 = 0.$$

Of these the first is necessary and sufficient for identifiability of the seven parameters and the remainder are genuine restrictions which specify the null hypothesis of independence. Note that we make use of $h_1(\theta) = 0$ in specifying the remaining restrictions; e.g. we take the simple form $h_2(\theta) = \theta_4 - \theta_1 \theta_2 = 0$ rather than the more direct form $h_2(\theta) = \theta_4 / \Sigma\theta_i - \theta_1 \theta_2 / (\Sigma\theta_i)^2 = 0$.

We suppose that a large random sample (with replacement) of n individuals yields n_1 in class a, n_2 in class b, ..., n_7 in class abc. Then

$$\log L(x, \theta) = \text{const} + \sum_{i=1}^{7} n_i \log \theta_i - n \log \left(\sum_{i=1}^{7} \theta_i \right),$$

and it follows that

$$\frac{\partial \log L(x, \theta)}{\partial \theta_i} = \frac{n_i}{\theta_i} - \frac{n}{\sum_{i=1}^{7} \theta_i} \quad (i = 1, 2, ..., 7),$$

and that

$$\mathbf{B}_\theta = \frac{1}{\Sigma\theta_i} \begin{bmatrix} \frac{1}{\theta_1} - \frac{1}{\Sigma\theta_i} & -\frac{1}{\Sigma\theta_i} & \cdots & -\frac{1}{\Sigma\theta_i} \\ -\frac{1}{\Sigma\theta_i} & \frac{1}{\theta_2} - \frac{1}{\Sigma\theta_i} & \cdots & -\frac{1}{\Sigma\theta_i} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{1}{\Sigma\theta_i} & -\frac{1}{\Sigma\theta_i} & \cdots & \frac{1}{\theta_7} - \frac{1}{\Sigma\theta_i} \end{bmatrix}.$$

Also

$$\mathbf{H}'_\theta = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -\theta_2 & -\theta_1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -\theta_3 & -\theta_2 & 0 & 1 & 0 & 0 \\ -\theta_3 & 0 & -\theta_1 & 0 & 0 & 1 & 0 \\ -\theta_2 \theta_3 & -\theta_3 \theta_1 & -\theta_1 \theta_2 & 0 & 0 & 0 & 1 \end{bmatrix},$$

from which $\mathbf{H}'_{1\theta} = [1\ 1\ 1\ 1\ 1\ 1\ 1]$, and, for any θ satisfying $h_1(\theta) = 0$,

$$\mathbf{B}_\theta + \mathbf{H}_{1\theta}\mathbf{H}'_{1\theta} = \text{diag}\left\{\frac{1}{\theta_1}, \frac{1}{\theta_2}, \ldots, \frac{1}{\theta_7}\right\}.$$

For the purpose of illustration we will suppose that $n = 100$, $n_1 = 46$, $n_2 = 24$, $n_3 = 7$, $n_4 = 15$, $n_5 = 3$, $n_6 = 4$ and $n_7 = 1$.

In this example unrestricted estimates are again easily obtained and Wald's test easily applied. We have

$$\theta^* = (\cdot46, \cdot24, \cdot07, \cdot15, \cdot03, \cdot04, \cdot01)$$

and $$\mathbf{h}'(\theta^*) = [0 \quad \cdot03960 \quad \cdot01320 \quad \cdot00780 \quad \cdot00227].$$

Also $$[\mathbf{H}'_{\theta^*}(\mathbf{B}_{\theta^*} + \mathbf{H}_{1\theta^*}\mathbf{H}'_{1\theta^*})^{-1}\mathbf{H}_{\theta^*}]^{-1} = \begin{bmatrix} 1{\cdot}04 & 0{\cdot}29 & -0{\cdot}10 & 0{\cdot}35 & 0{\cdot}98 \\ 0{\cdot}29 & 4{\cdot}57 & -0{\cdot}80 & -0{\cdot}27 & -1{\cdot}59 \\ -0{\cdot}10 & -0{\cdot}80 & 29{\cdot}68 & -3{\cdot}59 & -4{\cdot}77 \\ 0{\cdot}35 & -0{\cdot}27 & -3{\cdot}59 & 18{\cdot}58 & -5{\cdot}49 \\ 0{\cdot}98 & -1{\cdot}59 & -4{\cdot}77 & -5{\cdot}49 & 94{\cdot}24 \end{bmatrix}$$

so that Wald's statistic, $n\mathbf{h}'(\theta^*)[\mathbf{H}'_{\theta^*}(\mathbf{B}_{\theta^*} + \mathbf{H}_{1\theta^*}\mathbf{H}'_{1\theta^*})^{-1}\mathbf{H}_{\theta^*}]^{-1}\mathbf{h}(\theta^*)$, has value 1·1. Since this is less than the upper 5 per cent. point of a $\chi^2_{[4]}$ distribution, we accept the null hypothesis of independence.

We now turn to the problem of finding restricted estimates. We recall that to do so we have to obtain initial approximations $\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_7^{(1)}$ to them, invert the matrix

$$\begin{bmatrix} \mathbf{B}_{\theta^{(1)}} + \mathbf{H}_{1\theta^{(1)}}\mathbf{H}'_{1\theta^{(1)}} & -\mathbf{H}_{\theta^{(1)}} \\ -\mathbf{H}'_{\theta^{(1)}} & \mathbf{0} \end{bmatrix}$$

and use this inverse to obtain successive approximations to the restricted estimates. If we use as initial approximations the unrestricted estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_7$ and the suggested routine for matrix inversion (with $\mathbf{B}_{\theta^*} + \mathbf{H}_{1\theta^*}\mathbf{H}'_{1\theta^*}$ as B), then again by far the heaviest part of the computation has already been accomplished in applying the Wald test. Also, since $h_1(\theta^{(1)}) = 0$, this linear restriction will be satisfied at all subsequent stages and provides a useful partial check on the computations. We find that to two decimal places

$$\mathbf{Q}^{(1)\prime}_{\theta^*} = \begin{bmatrix} -0{\cdot}43 & -0{\cdot}21 & -0{\cdot}06 & -0{\cdot}20 & -0{\cdot}03 & -0{\cdot}06 & -0{\cdot}02 \\ 0{\cdot}35 & 0{\cdot}41 & -0{\cdot}05 & -0{\cdot}73 & 0{\cdot}02 & 0{\cdot}00 & 0{\cdot}01 \\ -0{\cdot}19 & 0{\cdot}40 & 0{\cdot}35 & 0{\cdot}13 & -0{\cdot}89 & 0{\cdot}15 & 0{\cdot}05 \\ 0{\cdot}37 & -0{\cdot}22 & 0{\cdot}47 & -0{\cdot}01 & 0{\cdot}10 & -0{\cdot}76 & 0{\cdot}05 \\ -0{\cdot}07 & 0{\cdot}24 & 0{\cdot}40 & 0{\cdot}09 & 0{\cdot}11 & 0{\cdot}18 & -0{\cdot}95 \end{bmatrix}$$

and

$$\mathbf{P}_{\theta^*}^{(1)} = \begin{bmatrix} 0{\cdot}21 & -0{\cdot}14 & -0{\cdot}04 & -0{\cdot}01 & -0{\cdot}02 & 0{\cdot}00 & 0{\cdot}00 \\ -0{\cdot}14 & 0{\cdot}14 & -0{\cdot}02 & 0{\cdot}03 & 0{\cdot}00 & -0{\cdot}02 & 0{\cdot}00 \\ -0{\cdot}04 & -0{\cdot}02 & 0{\cdot}04 & -0{\cdot}02 & 0{\cdot}01 & 0{\cdot}02 & 0{\cdot}00 \\ -0{\cdot}01 & 0{\cdot}03 & -0{\cdot}02 & 0{\cdot}01 & 0{\cdot}00 & -0{\cdot}01 & 0{\cdot}00 \\ -0{\cdot}02 & 0{\cdot}00 & 0{\cdot}01 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \\ 0{\cdot}00 & -0{\cdot}02 & 0{\cdot}02 & -0{\cdot}01 & 0{\cdot}00 & 0{\cdot}01 & 0{\cdot}00 \\ 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 & 0{\cdot}00 \end{bmatrix}.$$

The correction vector for our initial approximation θ^* to $\theta\dagger$ is

$$\frac{1}{n}\mathbf{P}_{\theta^*}^{(1)}\mathbf{D}\log L(x,\theta^*)+\mathbf{Q}_{\theta^*}^{(1)}\mathbf{h}(\theta^*)=\mathbf{Q}_{\theta^*}^{(1)}\mathbf{h}(\theta^*),$$

and so a second approximation $\theta^{(2)}$ to $\theta\dagger$ is $\theta^{(2)} = (\cdot47408, \cdot26035, \cdot07721, \cdot12294, \cdot02007, \cdot03646, \cdot00931)$. Also with $\mathbf{0}$ as our first approximation to $\hat{\boldsymbol{\lambda}}$ the corresponding correction vector is

$$\frac{1}{n}\mathbf{Q}_{\theta^*}^{(1)\prime}\mathbf{D}\log L(x,\theta^*)+\mathbf{R}_{\theta^*}^{(1)}\mathbf{h}(\theta^*)=\mathbf{R}_{\theta^*}^{(1)}\mathbf{h}(\theta^*)$$

and so $\boldsymbol{\lambda}^{(2)} = (-\cdot01512, -\cdot16470, -\cdot32127, -\cdot07438, -\cdot04517)$. On continuing this iterative process we eventually find that, to five decimal places,

$$\theta\dagger = (\cdot46903, \cdot26265, \cdot07824, \cdot12319, \cdot02055, \cdot03670, \cdot00964),$$
$$\hat{\boldsymbol{\lambda}} = (-\cdot03357, -\cdot18405, -\cdot42629, -\cdot05637, -\cdot00378).$$

Such accuracy is, of course, seldom required in practice.

It only remains to obtain a better estimate of the variance matrix of $\theta\dagger$ than is given by $\mathbf{P}_{\theta^*}^{(1)}/n$. This better estimate is $\mathbf{P}_{\theta\dagger}^{(1)}/n$ and either of the methods of section 4 may be used to obtain $\mathbf{P}_{\theta\dagger}^{(1)}$; the authors used the first method, finding

$$\mathbf{P}_{\theta\dagger}^{(1)} = \begin{bmatrix} \cdot21480 & -\cdot14271 & -\cdot03750 & -\cdot01052 & -\cdot01787 & -\cdot00078 & -\cdot00544 \\ -\cdot14271 & \cdot14298 & -\cdot01580 & \cdot02958 & \cdot00416 & -\cdot01858 & \cdot00037 \\ -\cdot03750 & -\cdot01580 & \cdot04253 & -\cdot01726 & \cdot00712 & \cdot01702 & \cdot00389 \\ -\cdot01052 & \cdot02958 & -\cdot01726 & \cdot01112 & -\cdot00274 & -\cdot00892 & -\cdot00126 \\ -\cdot01787 & \cdot00416 & \cdot00712 & -\cdot00274 & \cdot00674 & \cdot00194 & \cdot00066 \\ -\cdot00078 & -\cdot01858 & \cdot01702 & -\cdot00892 & \cdot00194 & \cdot00792 & \cdot00140 \\ -\cdot00544 & \cdot00037 & \cdot00389 & -\cdot00126 & \cdot00066 & \cdot00140 & \cdot00048 \end{bmatrix}.$$

Comparison of corresponding elements of $\mathbf{P}_{\theta^*}^{(1)}$ and $\mathbf{P}_{\theta\dagger}^{(1)}$ shows agreement to two decimal places and suggests that it might have been worthwhile to calculate $\mathbf{P}_{\theta^*}^{(1)}$ to more significant figures. Then $\mathbf{P}_{\theta^*}^{(1)}/n$ might have been used as an estimate of the variance matrix of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$, good enough for most practical purposes, and we

could have avoided the rather tedious final computation of $P^{(1)}_{\theta\dagger}$. In general, if an initial approximation $\theta^{(1)}$ to $\theta\dagger$ is fairly good, $P^{(1)}_{\theta^{(1)}}/n$ will provide a reasonable estimate of the variance matrix of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$.

Though strictly unnecessary, yet for the sake of completeness, we again calculate the value of the Lagrange-multiplier test statistic, which in any multinomial situation can be expressed as the usual

$$\sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}};$$

this value is 1·35, which is again smaller than the upper 5 per cent. point of a $\chi^2_{[4]}$ distribution, and again confirms the acceptance of $\mathscr{H}_0$. Also the statistic $2\log\Lambda$ of the likelihood-ratio test has value 1·29 and again we would accept $\mathscr{H}_0$.

7. Freedom Equation Specification of Restrictions

As we have noted above, the restrictions on the parameters $\theta_1, \theta_2, \ldots, \theta_s$ specifying $\mathscr{H}_0$ are quite often given in the form of freedom equations rather than of constraint equations. The null hypothesis then reads "the unknown parameters $\theta_1, \theta_2, \ldots, \theta_s$ can be expressed in the form

$$\theta_i = \theta_i(\alpha_1, \alpha_2, \ldots, \alpha_{s-r}) \quad (i = 1, 2, \ldots, s)", \tag{7.1}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_{s-r}$ are "freedom" parameters. To say that $\boldsymbol{\theta}$ can be expressed in this form is equivalent to imposing r constraints which, theoretically at least, can be found explicitly by elimination of $\boldsymbol{\alpha}$ from the equations (7.1). However, in most cases in which such a specification of $\mathscr{H}_0$ is given it is at best inconvenient to express the restrictions in the form of constraint equations, and the Lagrange-multiplier method, based as it is on restricted estimates, is much more easily adapted than is Wald's technique.

The natural approach in this case is to obtain restricted estimates as follows. We think of the likelihood function as a function of $\alpha_1, \alpha_2, \ldots, \alpha_{s-r}$ and obtain maximum-likelihood estimates $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{s-r}$, i.e. a value of $\boldsymbol{\alpha}$ which gives a maximum of the likelihood function when it is so regarded. The estimates $\hat{\theta}_i = \theta_i(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{s-r})$ $(i = 1, 2, \ldots, s)$ are then restricted estimates of the basic unknown parameters and the statistic $[\mathbf{D}\log L(x, \theta\dagger)]' \mathbf{B}^{-1}_{\theta\dagger}[\mathbf{D}\log L(x, \theta\dagger)]/n$ is, as before, distributed under $\mathscr{H}_0$ as $\chi^2_{[r]}$. Thus the Lagrange-multiplier test is very easily applied when $\mathbf{B}_\theta$ is non-singular. We see how this test applies in a familiar situation in the following example.

Example 4: quantal probit analysis. The probability that an individual will respond to a dose d_i of a stimulus is θ_i $(i = 1, 2, \ldots, s)$. A null hypothesis $\mathscr{H}_0$ states that θ_i can be expressed in the form $\theta_i = \Phi\{(d_i - \alpha_1)/\alpha_2\}$, where Φ is the standardized normal distribution function. In an experiment where n individuals are subjected to dose d_i, m_i respond and $n - m_i$ fail to respond $(i = 1, 2, \ldots, s)$.

By the usual probit analysis methods we obtain maximum-likelihood estimates $\hat{\alpha}_1, \hat{\alpha}_2$ of the freedom parameters α_1, α_2. These yield restricted maximum-likelihood estimates $\hat{\theta}_i = \Phi\{(d_i - \hat{\alpha}_1)/\hat{\alpha}_2\}$ $(i = 1, 2, \ldots, s)$ of the unknown probabilities associated with the different doses of the stimulus. We also have

$$\log L(x, \theta) = \text{const} + \sum_{i=1}^{s} [m_i \log \theta_i + (n - m_i)\log(1 - \theta_i)],$$

$$\frac{\partial \log L(x, \theta)}{\partial \theta_i} = \frac{m_i - n\theta_i}{\theta_i(1 - \theta_i)}$$

and $$\mathbf{B}_0^{-1} = \text{diag}\{\theta_1(1-\theta_1), \theta_2(1-\theta_2), \ldots, \theta_s(1-\theta_s)\},$$

so that the statistic on which the Lagrange-multiplier test of $\mathscr{H}_0$ is based is

$$\frac{1}{n}\sum_{i=1}^{s}\frac{(m_i - n\hat{\theta}_i)^2}{\hat{\theta}_i(1-\hat{\theta}_i)}.$$

Under $\mathscr{H}_0$, according to the foregoing general result, this is distributed, for large n, as $\chi^2_{[s-2]}$, a very familiar result.

If $\mathbf{B}_\theta$ is singular there are two possibilities. First, and most commonly, it may be necessary to impose additional (usually obvious) constraints on the θ's and hence on the α's in order to ensure identifiability of the parameters. For example, in investigating blood-groups (O, A, B, AB) with corresponding probabilities

$$\theta_1, \theta_2, \theta_3, \theta_4 \Big/ \sum_{i=1}^{4}\theta_i$$

we impose the identifiability restriction $\Sigma\theta_i = 1$ along with, say, the familiar freedom equation constraints $\theta_1 = \alpha_3^2$, $\theta_2 = \alpha_1^2 + 2\alpha_1\alpha_3$, $\theta_3 = \alpha_2^2 + 2\alpha_2\alpha_3$, $\theta_4 = 2\alpha_1\alpha_2$. The identifiability constraint, of course, implies the identifiability restriction $\sum_{i=1}^{3}\alpha_i = 1$ on the α's. Since the technique required in this situation is covered by that of the next section, we delay its consideration.

Secondly, certain of the constraints on θ implied by the freedom equations (7.1) may be necessary for identifiability of θ. The adaptation of the Lagrange-multiplier technique is then less elegant for it seems necessary to find explicitly from equations (7.1) constraints on θ which are necessary and sufficient for identifiability. If these are $h_1(\theta) = h_2(\theta) = \ldots = h_t(\theta) = 0$, and $\mathbf{H}_\theta$ is, as before, the $s \times t$ matrix $[\partial h_j(\theta)/\partial\theta_i]$, then the modified test statistic is

$$\frac{1}{n}[\mathbf{D}\log L(x, \theta\dagger)]'[\mathbf{B}_{\theta\dagger} + \mathbf{H}_{1\theta\dagger}\mathbf{H}'_{1\theta\dagger}]^{-1}[\mathbf{D}\log L(x, \theta\dagger)]$$

and is distributed as $\chi^2_{[r-t]}$ under $\mathscr{H}_0$.

8. Mixed Specification of Restrictions

It may indeed be natural, or convenient, to have mixed restrictions on θ, the null hypothesis being specified partly by constraint equations and partly by freedom equations; identifiability restrictions may also be necessary in this case. Let $\theta = (\psi, \phi)$ where $\psi = (\psi_1, \psi_2, \ldots, \psi_{s_1})$ and $\phi = (\phi_1, \phi_2, \ldots, \phi_{s_2})$, be the set of parameters under investigation. Suppose that the restrictions are expressed partly as constraint equations $h_1(\theta) = h_2(\theta) = \ldots = h_{r_1}(\theta) = 0$, of which the first t are necessary and sufficient for identifiability; the remaining $r_1 - t$ constraints, together with the restrictions specified by the freedom equations $\phi_i = \phi_i(\alpha_1, \alpha_2, \ldots, \alpha_{s_1 - r_2})$ $(i = 1, 2, \ldots, s_2)$, determine the null hypothesis $\mathscr{H}_0$.

In this case we again appeal to the Lagrange-multiplier method and hence, in the usual notation, employ the test statistic

$$\frac{1}{n}[\mathbf{D}\log L(x, \theta\dagger)]'[\mathbf{B}_{\theta\dagger} + \mathbf{H}_{1\theta\dagger}\mathbf{H}'_{1\theta\dagger}]^{-1}[\mathbf{D}\log L(x, \theta\dagger)],$$

where $\theta\dagger$ is the restricted estimate of θ. This statistic is, under $\mathscr{H}_0$, distributed as $\chi^2_{[r_1+r_2-t]}$. The restricted estimates $\theta\dagger$ are obtained by the following Lagrange-multiplier technique. Using $\phi_i = \phi_i(\alpha_1, \alpha_2, \ldots, \alpha_{s_1-r_1})$ $(i = 1, 2, \ldots, s_2)$, we consider the likelihood function as a function of ψ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{s_1-r_1})$, express the constraint equations in terms of ψ and α also, and then find by the method of section 6 the corresponding restricted estimates $\hat{\psi}$ and $\hat{\alpha}$ of ψ and α. Then $\theta\dagger = (\hat{\psi}, \hat{\phi})$ where

$$\hat{\phi}_i = \phi_i(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{s_1-r_1}) \quad (i = 1, 2, \ldots, s_2).$$

We now consider an application which makes use of this general technique.

Example 5. In a population the proportion of individuals possessing a certain attribute is known to be p. The probability $\phi(d)$ that an individual possesses the attribute depends on d, the measurement of a specified characteristic of the individual. The possible values of this characteristic are d_1, d_2, d_3, but the proportions in which these values occur in the population are unknown. We wish to test the hypothesis that the probability relation is of the form $\phi(d) = d/(d+\alpha)$. To achieve this we have available a large sample (with replacement) of n individuals; of the n_i in the sample who are found to have value d_i of the characteristic, m_i possess the attribute.

We let $\psi_i/\Sigma\psi_j$ be the proportion of individuals in the population having value d_i and write $\phi_i = \phi(d_i)$ $(i = 1, 2, 3)$. Our set of unknown parameters is then $\theta = (\psi, \phi)$. In forming our null hypothesis $\mathscr{H}_0$ we must remember that we know p and any model proposed must be tested for consistency with this fact. We impose immediately the necessary and sufficient identifiability restriction (as in Example 3)

$$h_1(\theta) = \sum_{i=1}^{3} \psi_i - 1 = 0;$$

the remaining restrictions then form our null hypothesis $\mathscr{H}_0$. The restriction imposed by our knowledge of p is

$$h_2(\theta) = \Sigma\psi_i\phi_i - p = 0,$$

and by the proposed probability relation we have the freedom equation restrictions

$$\phi_i = \frac{d_i}{d_i+\alpha} \quad (i = 1, 2, 3).$$

We have

$$\log L(x, \theta) = \text{const} + \Sigma n_i \log \psi_i - n \log \left(\sum_{i=1}^{3} \psi_i \right) + \sum_{i=1}^{3} \{ m_i \log \phi_i + (n_i - m_i) \log (1 - \phi_i) \}$$

and, when $h_1(\theta) = 0$, it is easy to show that

$$\mathbf{B}_\theta + \mathbf{H}_{1\theta}\mathbf{H}'_{1\theta} = \text{diag} \left\{ \frac{1}{\psi_1}, \frac{1}{\psi_2}, \frac{1}{\psi_3}, \frac{\psi_1}{\phi_1(1-\phi_1)}, \frac{\psi_2}{\phi_2(1-\phi_2)}, \frac{\psi_3}{\phi_3(1-\phi_3)} \right\}$$

and

$$\frac{\partial \log L(x, \theta)}{\partial \psi_i} = \frac{n_i - n\psi_i}{n\psi_i}, \frac{\partial \log L(x, \theta)}{\partial \phi_i} = \frac{m_i - n_i\phi_i}{\phi_i(1-\phi_i)}.$$

Hence the Lagrange-multiplier statistic can be expressed as

$$\sum_{i=1}^{3} \frac{(n_i - n\hat{\psi}_i)^2}{n\hat{\psi}_i} + \sum_{i=1}^{3} \frac{(m_i - n_i\hat{\phi}_i)^2}{n\hat{\psi}_i\hat{\phi}_i(1-\hat{\phi}_i)},$$

and, since in this case $r_1 = 2$, $r_2 = 2$ and $t = 1$, it is distributed as $\chi^2_{[3]}$ under $\mathscr{H}_0$.

To obtain $\hat{\psi}, \hat{\alpha}$ and hence $\theta\dagger = (\hat{\psi}, \hat{\phi})$ we use the usual iterative procedure with, when $\Sigma\psi_i - 1 = 0$,

$$\mathbf{H}'_{\psi,\alpha} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ \dfrac{d_1}{d_1+\alpha} & \dfrac{d_2}{d_2+\alpha} & \dfrac{d_3}{d_3+\alpha} & -\sum_{i=1}^{3} \dfrac{\psi_i d_i}{(d_i+\alpha)^2} \end{bmatrix}$$

$$\mathbf{H}'_{1\psi,\alpha} = [1 \quad 1 \quad 1 \quad 0]$$

$$\mathbf{B}_{\psi,\alpha} + \mathbf{H}_{1\psi,\alpha}\mathbf{H}'_{1\psi,\alpha} = \operatorname{diag}\left\{\frac{1}{\psi_1}, \frac{1}{\psi_2}, \frac{1}{\psi_3}, \frac{1}{\alpha}\sum_{i=1}^{3} \frac{\psi_i d_i}{(d_i+\alpha)^2}\right\}$$

$$\frac{1}{n}[\mathbf{D}\log L(x, \psi, \alpha)]' = \left[\frac{n_1}{n\psi_1} - 1 \quad \frac{n_2}{n\psi_2} - 1 \quad \frac{n_3}{n\psi_3} - 1 \quad -\frac{1}{n\alpha}\Sigma\left(m_i - \frac{n_i d_i}{d_i+\alpha}\right)\right].$$

To illustrate the theory we consider the case $p = 0{\cdot}545$, $d_1 = 1$, $d_2 = 2$, $d_3 = 4$, with the sample having $n = 200$, $n_1 = 53$, $n_2 = 103$, $n_3 = 44$, $m_1 = 21$, $m_2 = 54$, $m_3 = 23$. A convenient initial estimate of ψ is $\psi^{(1)} = (n_1/n, n_2/n, n_3/n) = (\cdot 265, \cdot 515, \cdot 220)$; on inspection $\alpha^{(1)} = 2{\cdot}0$ is seen to be a reasonable initial estimate of α; as usual we take $\lambda^{(1)} = (0, 0)$. We can quickly find by the matrix inversion method of section 4 that

$$\begin{bmatrix} \mathbf{B}_{\psi^{(1)},\alpha^{(1)}} + \mathbf{H}_{1\psi^{(1)},\alpha^{(1)}}\mathbf{H}'_{1\psi^{(1)},\alpha^{(1)}} & -\mathbf{H}_{\psi^{(1)},\alpha^{(1)}} \\ -\mathbf{H}'_{\psi^{(1)},\alpha^{(1)}} & \mathbf{0} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \cdot 188 & -\cdot 136 & -\cdot 052 & -\cdot 338 & -\cdot 348 & \cdot 169 \\ -\cdot 136 & \cdot 250 & -\cdot 114 & \cdot 031 & -\cdot 507 & -\cdot 015 \\ -\cdot 052 & -\cdot 114 & \cdot 166 & \cdot 307 & -\cdot 145 & -\cdot 153 \\ -\cdot 338 & \cdot 031 & \cdot 307 & \cdot 909 & -3{\cdot}940 & 7{\cdot}999 \\ -\cdot 348 & -\cdot 507 & -\cdot 145 & -3{\cdot}940 & -1{\cdot}969 & 1{\cdot}968 \\ \cdot 169 & -\cdot 015 & -\cdot 153 & 7{\cdot}999 & 1{\cdot}968 & -3{\cdot}996 \end{bmatrix}$$

Also $$\frac{1}{n}[\mathbf{D}\log L(x, \psi^{(1)}, \alpha^{(1)})]' = [0 \quad 0 \quad 0 \quad 0{\cdot}0012]$$

and $$\mathbf{h}'(\psi^{(1)}, \alpha^{(1)}) = [0 \quad -0{\cdot}0525],$$

so that, using the iterative method of section 6, we have

$$\psi^{(2)} = (\cdot 256, \quad \cdot 516, \quad \cdot 228); \qquad \alpha^{(2)} = 1{\cdot}581; \qquad \lambda^{(2)} = (-\cdot 108, \quad \cdot 219).$$

If we continue this iterative process we find that, correct to three decimal places,

$$\hat{\psi} = (\cdot 256, \quad \cdot 516, \quad \cdot 228); \qquad \hat{\alpha} = 1{\cdot}621; \qquad \hat{\lambda} = (-\cdot 121, \quad \cdot 222);$$

so that the value of the Lagrange-multiplier test statistic is 7·88. This is greater than the upper 5 per cent. point of a $\chi^2_{[3]}$ distribution, so that in this case we reject $\mathcal{H}_0$, and there is no need to proceed further.

It is interesting to compare this with an alternative procedure. We might carry out a test of $\mathscr{H}_0$ in two stages: first testing the hypothesis $\mathscr{H}_{01}$ that $\phi_i = d_i/(d_i+\alpha)$ $(i = 1, 2, 3)$ by a Lagrange-multiplier $\chi^2_{[2]}$ test, and then $\mathscr{H}_{02}$ that $h_2(\theta) = 0$ by a Wald $\chi^2_{[1]}$ test. The first test yields a non-significant value 5·28 and so we would accept $\mathscr{H}_{01}$; the second test, also, yields a non-significant value 2·42, and so we would accept $\mathscr{H}_{02}$. This procedure would thus result in the acceptance of $\mathscr{H}_0$. The reason for the discrepancy with the result of the preceding test is that here we are subjecting the model to two partial tests which separately are not sufficient to result in the rejection of $\mathscr{H}_0$, whereas the former subjects it to a single complete test of the hypothesis $\mathscr{H}_0$.

REFERENCES

AITCHISON, J. & SILVEY, S. D. (1958), "Maximum-likelihood estimation of parameters subject to restraints", *Ann. Math. Statist.*, 29, 813–828.

SILVEY, S. D. (1959), "The Lagrange-multiplier test", *Ann. Math. Statist.*, 30, 389–407.

WALD, A. (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large", *Trans. Amer. Math. Soc.*, 54, 426–482.

4 MULTIPLE HYPOTHESIS TESTING

4.1 Introduction

The developments described in §3 refer to the testing of a single hypothesis within a model. In many practical situations there may be more than one hypothesis of interest, and sometimes a large number of such hypotheses. In linear modelling it is precisely towards such multiple hypothesis testing that the package of techniques known as analysis of variance and analysis of covariance is directed. Similarly in multivariate normal modelling there are corresponding methodologies such as analysis of dispersion for dealing with multiple hypothesis problems. Where the parametric modelling does not fall into these special categories the natural follow-up question to the successful development of methods for single hypothesis investigation is whether it is possible to develop methodology appropriate to the problems of multiple hypotheses. The purpose of the three papers of Aitchison (5:1962, 6:1964, 7:1965) was to answer this question in the affirmative.

First, it should be emphasised that the only non-controversial solution to multiple hypothesis testing is probably through a full decision-theoretic approach. In particular applications it is seldom, at least in the author's experience, that the necessary loss or utility structure is available. This is, of course, as true of the essentially ad hoc practices adopted in linear and multivariate normal theory as of the general parametric case. Our concern must be to ensure that the ad hoc procedures developed are sensible, and to achieve this, counterparts of such linear model

concepts as nesting, orthogonality, partitioning and simultaneity have to be found for the general parametric model. Three basic concepts, restricted parametric tests, separable hypotheses and confidence-region tests, are developed in the three papers and we now outline the role these play in supplying the necessary methodology.

4.2 Large-sample restricted parametric tests

If in constructing a test of ω_2 within a model $\omega_0 = \Theta$ we take the alternative to be $\omega_1 - \omega_2$, where $\omega_2 \subset \omega_1 \subset \omega_0$ instead of the full and natural alternative $\omega_0 - \omega_2$, we say that we have a restricted (generalised likelihood ratio) test of ω_2 against $\omega_1 - \omega_2$ within ω_0. After showing how to construct critical regions and power functions for various general parametric specifications of ω_0, ω_1 and ω_2 Aitchison (5:1962) then presents the following developments.

1. A simple relation $T_{21} = T_{20} - T_{10}$ is obtained between test statistics, where T_{ij} is any large-sample equivalent test statistic for testing ω_i against $\omega_j - \omega_i$. This can be used as a computational device when T_{21} is difficult to evaluate directly.
2. The effect of restriction of the alternative hypothesis on the increase in power is quantified.
3. For nested hypotheses a systematic form of investigation, analysis of discrepancy, is devised with a systematic tabular form of computation similar to analysis of variance.

4.3 Separable hypotheses

If interest is in two hypotheses ω_1 and ω_2 which are not nested, there may be some doubt as to which of the nested sets, $(\omega_0, \omega_1, \omega_1 \cap \omega_2)$ or $(\omega_0, \omega_2, \omega_1 \cap \omega_2)$, to investigate. In such

circumstances it is obviously advisable to examine both sets, being particularly careful about the interpretation of the results. There can arise, however, an important relationship between the two hypotheses which makes the double inspection unnecessary and so eases the interpretation. This occurs when the restricted tests of $\omega_1 \cap \omega_2$ against $\omega_2 - \omega_1 \cap \omega_2$ and of $\omega_1 \cap \omega_2$ against $\omega_1 - \omega_1 \cap \omega_2$ use the same critical regions as the unrestricted tests of ω_1 against $\omega_0 - \omega_1$ and of ω_2 against $\omega_0 - \omega_2$ respectively, so that examination of the two nested sets involves the application of the same three unrestricted tests, of ω_1 against $\omega_0 - \omega_1$, of ω_2 against $\omega_0 - \omega_2$, and of $\omega_1 \cap \omega_2$ against $\omega_0 - \omega_1 \cap \omega_2$. Aitchison (5:1962) terms such hypotheses separable with resepct to the type of test used, points out the analogy with orthogonality and devises a simple criterion, in terms of the H_1 and H_2 matrices associated with ω_1 and ω_2 and the information matrix B, for recognising when two hypotheses are separable. It is then shown that when hypotheses are separable it is possible to partition the statistic associated with the testing of their intersection into a sum of two statistics appropriate to the testing of the hypotheses separately. Thus for the general parametric case there is a counterpart to linear model partitioning, by which we can follow up the rejecting of the intersection hypothesis by investigation of the larger separable hypotheses.

A recent review with some developments of the concept of separability is to be found in Mathieu (1978).

4.4 Confidence-region tests

The motivation underlying the concept of confidence region tests introduced in Aitchison (6:1964) and further developed in Aitchison (7:1965) is to try to place some overall control on

significance level in multiple hypothesis testing. The rationale is based on the following important points.

1. In multiple hypothesis situations the concept of significance level after the first test has been applied has no real meaning, and is little more than an arbitrary set of rules.
2. In hypothesis testing the rejection of a hypothesis is a much more positive act than its 'acceptance', which may be merely due to weakness in bringing enough evidence to bear against the hypothesis. Thus any testing procedure which proceeds by investigating hypotheses larger (regarded as subsets of the parameter set) and containing an already rejected hypothesis is to be preferred to one which moves from an accepted hypothesis to the examination of proper subsets of that hypothesis.
3. Much of hypothesis-testing is a search for an appropriate model as a basis for further activity, such as estimation or prediction. Thus acceptance of a hypothesis is in a sense a license to introduce a higher-dimensional parameter. It seems reasonable therefore to start with 'low-parameter' hypotheses, rejection being the requirement for moving to a 'higher parameter' hypothesis. In other words, things are simple or simply random until we have real evidence to believe otherwise. This is in accord with one of the principles of inference propounded by Jeffreys (1961,p.47).

All this suggests that we should look for a system of testing controlled by a single 'significance level', which proceeds from the smallest hypothesis through to the largest only by the process of rejection. The method of Aitchison (6:1964) provides such a

tool and is best explained initially in relation to a nested sequence of hypotheses $\omega_k \subset \omega_{k-1} \subset \ldots \subset \omega_1 \subset \omega_0 = \Theta$. It is shown that a confidence region $C(x)$, with confidence coefficient $1-\alpha$, can be constructed in such a way that it can be used for testing ω_k within Θ at significance level α instead of a critical region $R \subset X$, in the sense that

$$\omega_k \cap C(x) = \emptyset \quad \text{iff } x \in R.$$

With this confidence region $C(x)$, based on appropriately testing the smallest hypothesis ω_k, we can proceed to test other hypotheses, rejecting ω_i $(i < k)$ if

$$\omega_i \cap C(x) = \emptyset,$$

that is if ω_i has no parameter points in common with the confidence region originally set up for testing ω_k. Thus testing is under the control of the single significance level α, or equivalently the confidence coefficient $1-\alpha$. For a non-nested collection of hypotheses the basic hypothesis on which $C(x)$ is built should be the intersection hypothesis. Then testing proceeds in exactly the same way as set out above.

Having established the principles of the confidence-region testing of multiple hypotheses, Aitchison (6:1964) then proceeds to turn the above criterion for rejection into an easily operable system for linear models and the general parametric model. Only some of the more important aspects need be indicated here.

1. The test procedure is shown to be equivalent to use of the standard (F for linear; chi-squared for general parametric) test statistic for testing ω_i within ω_0 but against a critical

value larger than that associated with significance level α. In other words the standard test statistic is used but at a smaller, sometimes much smaller, 'equivalent significance level' than α. These equivalent significance levels are shown to be easily computed. Thus if ω_k is rejected at level α the procedure makes it increasingly more difficult to reject larger hypotheses.

2. The method is shown to agree with, and indeed to be a generalisation of, the Scheffé (1953) S-method for the linear model, and of the Goodman (1964) method of multiple comparison in contingency tables.

3. The practical implications for such activities as data snooping, introspection of an accepted hypothesis and enlarging rejected hypotheses are discussed with examples.

4. It is shown that a disturbing feature may arise in standard practice in linear models, that rejection of $\omega_1 \cap \omega_2$, if ω_1 and ω_2 are orthogonal, automatically means the rejection of at least one of the wider hypotheses ω_1 and ω_2. If there are a priori grounds for believing that rejection of ω can only arise through the rejection of ω_1 or ω_2 or both, then we cannot object to this feature, although we may ask if the purpose of the initial test of ω has been clearly thought out. Often, however, there are no such a priori grounds and indeed the testing of ω_1 and ω_2 may arise as an afterthought to the rejection of ω. How sensible is it then to use the above procedure in which enlargement of ω almost automatically follows its rejection? Confidence region testing protects against this disturbing feature. Rejection of ω does not now imply automatic or nearly automatic rejection of one or other

of ω_1 and ω_2, and again we have the feature, desirable in many practical situations, of much more stringent assessment of data when larger hypotheses are under consideration.

In Aitchison (6:1964) although appropriate confidence regions for the linear and general parametric model are obtained no general principle for their construction is considered. In Aitchison (7:1965) such a principle of construction is presented, relating the confidence-region tests to generalised likelihood ratio tests. Specifically if $\Lambda(x;a)$ is the generalised likelihood ratio test statistic for testing the hypothesis $\omega(a) = \{\theta : h(\theta) = a\}$ with critical value $c_\alpha(\Lambda)$ for all a then

$$C(x) = [\theta : \Lambda\{x;h(\theta)\} \leqq c_\alpha(\Lambda)]$$

is a suitable basic confidence region for investigating multiple hypotheses containing $\omega(a)$. In particular it is shown that the confidence-region method uses the usual generalised likelihood ratio test statistic for testing ω_i within the model, but at a more stringent significance level. It is thus shown that the method is equivalent to fixed-odds likelihood-ratio testing, which has been separately advocated as a simultaneous test procedure by Gabriel (1964).

This availability of a principle of construction allows Aitchison (7:1965) to treat a multivariate normal problem in an exact form rather than through the asymptotic theory associated with the general parametric form.

4.5 Discussion

The developments described in this and the previous section, particularly for the general parametric model, provide a set of

powerful tools for the facing of new problems. It should perhaps be emphasised that just as their strength is their generality so their weakness may be that they are rougher tools than more precise ones devisable through the particularity of the problem. It should always be the duty of the statistician to produce as exact methods as possible for his particular parametric situation and only when this proves intractable should recourse be made to the general tools. Their usefulness continues to lie in the undoubted fact that intractability is a persistent intruder in the real world of practical problems.

AITCHISON, J. (1962)

Large-sample restricted parametric tests

Reprinted from *J. R. Statist. Soc.* B24, 234-50

Large-sample Restricted Parametric Tests

By JOHN AITCHISON

University of Glasgow

[Received November 1961]*

SUMMARY

Restricted tests are designed to examine a hypothesis not against the natural complete alternative but against some subhypothesis of that alternative. The theory of such tests provides a unification of many aspects of large-sample tests—the analysis of nested hypotheses, the possibility of using tests of increased power and of easing certain computational procedures, the notions of separable hypotheses and of partitioning a test-statistic. While many standard results can thus be brought within the framework of this theory the emphasis throughout the paper is on providing the consulting statistician with manageable tools for non-standard situations.

1. INTRODUCTION

To define a restricted parametric test we require the following setting. An experiment is described by a k-dimensional random variable X whose unknown distribution belongs to a known class of distributions, indexed by an s-dimensional parameter θ. The set of possible parameters is a given subset ω_0 of s-dimensional Euclidean space R^s, and the restriction of θ to ω_0 leaves the s components of θ independent. We denote by θ_0 the true parameter value associated with X and assume that X is either discrete or continuous so that the class of distributions can be specified by a class of elementary probability laws $\{f(\cdot, \theta) : \theta \in \omega_0\}$. We can thus refer to a distribution with probability law $f(\cdot, \theta)$ simply by its label θ. Also if ω is a proper subset of ω_0 (written $\omega_0 \supset \omega$) it is convenient to refer to the hypothesis that $\theta_0 \in \omega$ shortly as the hypothesis ω or simply as ω; a basic hypothesis, here ω_0, assumed to be true we call the model ω_0. A set x of n independent observations on the random variable X is available as the basis of any test we may consider.

Suppose that $\omega_0 \supset \omega_1 \supset \omega_2$ and that we wish to test the null hypothesis ω_2. If we construct a test of ω_2 restricting the alternative of interest to $\omega_1 - \omega_2$ rather than to the complete alternative $\omega_0 - \omega_2$ the resulting test is termed a *restricted* test of ω_2 against $\omega_1 - \omega_2$ within the model ω_0 to distinguish it from an unrestricted test of ω_2 against $\omega_0 - \omega_2$. For example, if, under ω_0, the likelihood of θ for given x is

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

then a test which uses a critical region of the form

$$\left\{x : \sup_{\theta \in \omega_1} L(x, \theta) / \sup_{\theta \in \omega_2} L(x, \theta) > c\right\}$$

is a restricted generalized likelihood ratio test of ω_2 against $\omega_1-\omega_2$ within ω_0 whereas the unrestricted generalized likelihood ratio test would use a critical region of the form

$$\left\{x : \sup_{\theta\in\omega_0} L(x,\theta)/\sup_{\theta\in\omega_2} L(x,\theta) > c\right\}.$$

To construct such restricted tests we could, of course, apply the available theory of unrestricted tests to obtain an unrestricted test of ω_2 against all alternatives within the adopted model ω_1. Thus, to provide proper motivation for the development of a theory of restricted tests we must not only respect the customary demand to demonstrate usefulness, but must also examine why and where the available theory of unrestricted tests is deficient. These two aspects we now consider in turn.

2. Uses of Restricted Tests

2.1. *Nested Hypotheses*

The experimenter may be fortunate enough to have thought out his theory sufficiently clearly and designed his experiment so well that he can frame his problem as one involving a set of nested hypotheses $\omega_1, \omega_2, \ldots, \omega_m$ within the basic model ω_0 in the sense that $\omega_0 \supset \omega_1 \supset \omega_2 \ldots \supset \omega_m$. He then has to decide how far along this path of specialization the results of his experiment will allow him to go. One rational method of procedure for this essentially multiple decision problem is the following.

(i) First test ω_1 against $\omega_0-\omega_1$. If ω_1 is accepted proceed to the next step.

(ii) Test ω_2 within the accepted model ω_1; that is, carry out a restricted test of ω_2 against $\omega_1-\omega_2$ within the model ω_0. Also, as a precaution against proceeding along the path of specialization by near significant steps, carry out an unrestricted test of ω_2 against $\omega_0-\omega_2$. Only if both tests yield non-significant results accept ω_2 and proceed to the next step.

(iii) Carry out restricted tests of ω_3 against $\omega_2-\omega_3$ and against $\omega_1-\omega_3$, and an unrestricted test of ω_3 against $\omega_0-\omega_3$. If all three tests produce non-significant results accept ω_3 and proceed to the next obvious step.

Thus at each stage of this analysis after the first the experimenter meets the need for restricted tests.

It is clear that we have avoided the difficulties of the original multiple decision problem only by tacitly neglecting the problems associated with the multiple significant tests in the above analysis. In the absence of any really satisfactory theory for either problem we have adopted the above practical procedure because it has a certain intuitive appeal. A fuller investigation of optimum properties of the suggested procedure, possibly in relation to realistic loss functions, seems desirable but is not considered in this paper. Such nested hypotheses and similar analyses occur in standard statistical problems, particularly in analyses of variance. For example, if ω_0 is the model associated with a fixed-effects two-way classification design with n replicates in each of the (A_i, B_j) treatment combinations then ω_1 might be the hypothesis of no interaction and ω_2 the hypothesis of no interaction and no A-treatment effect. Our aim in this paper is to provide a course of action in non-standard situations, and to illustrate this we shall make repeated use of the following simple examples.

Example 1. The experimenter knows that X is $N(\theta_1, \theta_2^2)$ with $\omega_0 = \{\theta : \theta_2 > 0\}$, and is interested in testing the specializations $\omega_1 = \{\theta\in\omega_0 : \theta_1-\theta_2 = 0\}$, the hypothesis of equality of mean and standard deviation, and $\omega_2 = \{\theta\in\omega_1 : \theta_1 = 1\}$, the simple

hypothesis that X is $N(1,1)$. Since $\omega_0 \supset \omega_1 \supset \omega_2$ the problem involves the testing of nested hypotheses.

Example 2. The experimenter is dealing with k independent normal distributions and his successive specializations are ω_1, the hypothesis of equality of coefficients of variation, that is of proportionality of standard deviation to mean, and ω_2, the hypothesis that the k coefficients of variation are equal to a specified constant c. Here $X = (X_1, X_2, \ldots, X_k)$ with independent components, X_i being $N(\theta_{2i-1}, \theta_{2i}^2)$ $(i = 1, 2, \ldots, k)$. The parameter θ thus has $2k$ components and the model and nested hypotheses of interest are

$$\omega_0 = \{\theta : \theta_{2i} > 0 \quad (i = 1, 2, \ldots, k)\},$$

$$\omega_1 = \{\theta \in \omega_0 : \theta_2/\theta_1 = \theta_4/\theta_3 = \ldots = \theta_{2k}/\theta_{2k-1}\},$$

$$\omega_2 = \{\theta \in \omega_1 : \theta_2/\theta_1 = c\}.$$

Example 3. The experimenter is concerned with the emission rates of particles from p independent radioactive specimens, but can count the number of particles emitted from a specimen only together with those arising from an independent background "noise". His experiment consists of $p+1$ independent counts over equal intervals of time from the following sources—(i) background noise alone (described by a discrete random variable X_1), (ii) the first specimen and background noise (described by X_2), (iii) the second specimen and background noise (described by X_3), and so on. This experiment, described by $X = (X_1, X_2, \ldots, X_{p+1})$, is replicated n times. The hypotheses of interest are ω_1, that the p specimens have been arranged in sequence, the ith containing i units of radioactivity; ω_2, that the unit of radioactivity in ω_1 corresponds to a specified emission rate; and ω_3, that ω_2 holds with a specified emission rate for background noise. If a Poisson type model is a satisfactory description of this experiment then the independent components of X are such that X_i is distributed as a Poisson random variable with mean parameter θ_i. Then

$$\omega_0 = \{\theta : \theta_i > 0 \quad (i = 1, 2, \ldots, p+1)\},$$

and the nested hypotheses are

$$\omega_1 = \{\theta \in \omega_0 : \theta_2 - \theta_1 = \theta_3 - \theta_2 = \ldots = \theta_{p+1} - \theta_p\},$$

$$\omega_2 = \{\theta \in \omega_1 : \theta_2 - \theta_1 = d\},$$

$$\omega_3 = \{\theta \in \omega_2 : \theta_1 = a\}.$$

Example 4. The individuals in each of k independent populations possess either an attribute A or B or both A and B (denoted by AB). The random sampling with replacement of n individuals from each of these populations can thus be regarded as n replicates of an experiment described by a random variable $X = (X_1, X_2, \ldots, X_k)$ whose components are independent, X_i having a trinomial distribution with class probabilities $(\theta_{i1}, \theta_{i2}, \theta_{i3})/(\theta_{i1} + \theta_{i2} + \theta_{i3})$ $(i = 1, 2, \ldots, k)$. Here $\omega_0 = \{\theta : \text{all } \theta_{ij} > 0\}$. The first hypothesis ω_1 to be considered is that the attributes are independent in each population in the sense that the proportion of individuals with AB in a population is the product of the proportions with A and with B separately, that is

$$\omega_1 = \left\{\theta \in \omega_0 : \frac{\theta_{i3}}{\theta_{i1} + \theta_{i2} + \theta_{i3}} = \frac{\theta_{i1}\theta_{i2}}{(\theta_{i1} + \theta_{i2} + \theta_{i3})^2} \quad (i = 1, 2, \ldots, k)\right\},$$

to be followed by consideration of the nested hypotheses

$$\omega_2 = \left\{\theta \in \omega_1 : \frac{\theta_{ij}}{\theta_{i1}+\theta_{i2}+\theta_{i3}} = \frac{\theta_{1j}}{\theta_{i1}+\theta_{i2}+\theta_{i3}} \quad (i = 1,2,\ldots,k-1; \quad j = 1,2)\right\},$$

the hypothesis that in addition to independence the k populations possess the attributes in the same proportions; and

$$\omega_3 = \left\{\theta \in \omega_2 : \frac{\theta_{11}}{\theta_{11}+\theta_{12}+\theta_{13}} = \frac{1}{2}\right\},$$

the hypothesis that in each population the proportions in the three categories are $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$. This is a situation where, of course, it is necessary to impose identifiability conditions, for instance

$$\sum_{j=1}^{3} \theta_{ij} = 1 \quad (i = 1,2,\ldots,k).$$

2.2. *Increasing Power by Restricting Alternatives*

While the experimenter is aware that all alternatives to ω_2 in $\omega_0 - \omega_2$ are possible, he may be especially concerned about a certain subset of these alternatives, say $\omega_1 - \omega_2$, where $\omega_0 \supset \omega_1 \supset \omega_2$. This subset may contain those alternatives which the experimenter considers most likely or most important or whose rejection, when true, are expected to prove most costly. He may then decide, in constructing his test of ω_2 against $\omega_0 - \omega_2$, to concentrate attention on those alternatives in $\omega_1 - \omega_2$ in the hope that the resulting restricted test will have appreciably greater power for these alternatives than an unrestricted test constructed against the wider class of alternatives. By so doing, of course, he allows the alternatives in $\omega_0 - \omega_1$ to fend for themselves. The idea is an old one. For example, the familiar one-sided t-test is designed to meet the need for a test of the hypothesis that a normal distribution mean $\theta_0 = c$ against the restricted alternative that $\theta_0 > c$, say, as opposed to the set of possible alternatives that $\theta_0 \neq c$.

The technique has been considerably exploited as a means of improving the chi-squared test of goodness-of-fit. Recent work on this subject, together with a short survey, has been undertaken by Fix, Hodges and Lehmann (1959), who introduced the term restricted chi-squared test. The device has, however, a wider application than in their particular multinomial situation; indeed we shall see that the theory presented later includes these restricted chi-squared tests as special cases.

Examples of non-standard situations are again given by reference to our simple illustrations. In Example 1 the experimenter may wish to test the hypothesis that X is $N(1, 1)$ and have strong grounds for believing that in any alternative the mean and standard deviation of the underlying normal distribution are equal. In such circumstances a restricted test of ω_2 against $\omega_1 - \omega_2$ within ω_0 is called for. Similar considerations also apply to Examples 2–4. For instance, in Example 3, in constructing a test of the hypothesis ω_2 that the emission rates for the *specimens* are $d, 2d, \ldots, pd$, the experimenter may be fairly sure that these emission rates are proportional to the first p integers, that is that ω_1 holds, and so will want a restricted test of ω_2 against $\omega_1 - \omega_2$ within ω_0.

2.3. *Easing the Evaluation of a Test Statistic*

For an unrestricted test of a hypothesis ω_2 against all alternatives within a basic model ω_1 it may well prove a computational advantage to introduce a more basic

model ω_0 and to construct a restricted test of ω_2 against $\omega_1 - \omega_2$ within the wider setting of this model. The reasons for this will be seen early in the next section, where we examine the computational difficulties that may arise in a direct application of unrestricted tests.

2.4. *Separable Hypotheses*

The experimenter may be interested in two hypotheses ω_1 and ω_2 which are not nested, and so may be in some doubt as to which of the nested sets, $(\omega_0, \omega_1, \omega_1 \cap \omega_2)$ or $(\omega_0, \omega_2, \omega_1 \cap \omega_2)$, to investigate. In such circumstances he would be well advised to examine both sets, being particularly careful about the interpretation of the results. There can arise, however, an important relationship between the two hypotheses which makes the double inspection unnecessary and so eases the interpretation. This occurs when the restricted tests of $\omega_1 \cap \omega_2$ against $\omega_2 - \omega_1 \cap \omega_2$ and of $\omega_1 \cap \omega_2$ against $\omega_1 - \omega_1 \cap \omega_2$ use the same critical regions as the unrestricted tests of ω_1 against $\omega_0 - \omega_1$ and of ω_2 against $\omega_0 - \omega_2$ respectively, so that examination of the two nested sets involves the application of the same three unrestricted tests, of ω_1 against $\omega_0 - \omega_1$, of ω_2 against $\omega_0 - \omega_2$, and of $\omega_1 \cap \omega_2$ against $\omega_0 - \omega_1 \cap \omega_2$. We shall term such hypotheses *separable* with respect to the type of test used. Such a notion occurs in analysis of variance where, for example, orthogonal designs are aimed precisely at such separation of hypotheses. It is important to be able to recognize easily when two hypotheses are separable and we shall see later that this can be done by the application of a simple criterion provided by the theory of restricted tests.

Example 5. The three dimensions of cuboids produced by a certain process are described by a random variable $X = (X_1, X_2, X_3)$, with independent components and where X_i has probability density $x^{p-1} \exp(-x/\theta)/\{\theta^p \Gamma(p)\}$ $(x > 0)$ with known p. Here $\omega_0 = \{\theta : \theta_i > 0 \quad (i = 1, 2, 3)\}$. If the experimenter wishes to test $\omega_1 = \{\theta \in \omega_0 : p^3 \theta_1 \theta_2 \theta_3 = a^3\}$, the hypothesis that the average volume is a^3 and $\omega_2 = \{\theta \in \omega_0 : \theta_1 = \theta_2 = \theta_3\}$, the hypothesis that the three mean dimensions are equal, then the question of the separability of ω_1 and ω_2 may well arise.

Example 6. A random sample with replacement of n individuals is taken from a genetic population whose individuals belong to one or other of the three types—dominant, hybrid, recessive. We thus have n independent observations on a trinomial random variable X with probabilities $(\theta_1, \theta_2, \theta_3)/(\theta_1 + \theta_2 + \theta_3)$ for dominant, hybrid and recessive types respectively, and $\omega_0 = \{\theta : \theta_i > 0 \quad (i = 1, 2, 3)\}$. The experimenter wishes to study $\omega_1 = \{\theta \in \omega_0 : [\theta_1/(\theta_1 + \theta_2 + \theta_3)]^{\frac{1}{2}} + [\theta_3/(\theta_1 + \theta_2 + \theta_3)]^{\frac{1}{2}} = 1\}$, the hypothesis that the population is genetically stable, and $\omega_2 = \{\theta \in \omega_0 : \theta_1 = \theta_3\}$, the hypothesis of equal proportions of dominants and recessives, and so the question of the separability of these two hypotheses may arise. In this case we must again impose an identifiability condition $\theta_1 + \theta_2 + \theta_3 = 1$.

2.5. *Partitioning a Test Statistic*

The theory of restricted tests is also useful in the situation where the experimenter has found a hypothesis ω_3 unacceptable and has turned his attention to an examination of components ω_1 and ω_2 such that $\omega_3 = \omega_1 \cap \omega_2$. We have then to investigate the possibility of partitioning the test statistic for ω_3 against $\omega_0 - \omega_3$ into components for testing ω_1 against $\omega_0 - \omega_1$ and ω_2 against $\omega_0 - \omega_2$. We shall see that such partitioning is possible when the hypotheses ω_1 and ω_2 are separable. The familiar partitioning of a chi-squared goodness-of-fit test statistic arises as a particular case of the general theory which indeed in this multinomial case provides an extension of the results as

usually presented. Illustrations of the uses of partitioning can be provided by Examples 5 and 6. In Example 5 if the experimenter has found

$$\omega_3 = \left\{\theta \in \omega_0 : \theta_1 = \theta_2 = \theta_3 = \frac{a}{p}\right\}$$

unacceptable he may wish to investigate the composition $\omega_3 = \omega_1 \cap \omega_2$ where ω_1 and ω_2 are as previously defined. Again in Example 6 the rejection of $\omega_3 = \{\theta \in \omega_0 : \theta_1 = \theta_3 = \frac{1}{4},\ \theta_2 = \frac{1}{2}\}$ may be followed by examination of ω_1 and ω_2 since $\omega_3 = \omega_1 \cap \omega_2$.

3. Difficulties in Applying Available Unrestricted Tests

To clarify subsequent discussion and to establish notation we first recall briefly the available large sample unrestricted tests as developed by Wald (1943), Rao (1948), Aitchison and Silvey (1958, 1960), and Silvey (1959).

3.1. *Assumptions and Notation*

We assume that all functions and sets introduced satisfy certain regularity conditions, such as differentiability, compactness, etc., required for the results of this and subsequent sections to hold. To attempt to introduce these conditions, almost invariably satisfied in practice, in order to provide a fully rigorous treatment would cause considerable embarrassment in a paper which seeks to alleviate the lot of the experimenter and consulting statistician. For convenience we drop from the notation the dependence on sample size n; this should cause no misunderstanding in the context of this paper. What must be emphasized is that all the properties we investigate are asymptotic, valid for large n, and from now on we suppose that n is always large.

3.2. *The Likelihood and Associated Constructs*

The likelihood of $\theta(\in\omega_0)$, given x, is denoted by $L(x, \theta)$. We suppose that a maximum likelihood estimator θ^i of θ under any hypothesis ω_i considered exists, that is a function θ^i from the sample space to R^s can be defined for each n such that for every x we have $\theta^i(x) \in \omega_i$ and

$$L(x, \theta^i(x)) = \sup_{\theta \in \omega_i} L(x, \theta).$$

We denote by $\mathbf{D}\log L(x, \theta)$ the $s \times 1$ column vector whose ith component is $\partial \log L(x, \theta)/\partial\theta_i$, and by $\mathbf{B}_\theta$ the $s \times s$ information matrix based on ω_0 whose (i,j)th component is $-(1/n)E_\theta(\partial^2 \log L(\cdot, \theta)/\partial\theta_i \partial\theta_j)$, where E_θ denotes expectation with respect to the distribution with elementary probability law $L(\cdot, \theta)$. The matrix $\mathbf{B}_\theta$ is assumed non-singular for all $\theta \in \omega_0$; the minor adjustments necessary to the theory in the contrary case are indicated in section 3.9.

3.3. *Hypotheses and Associated Constructs*

We confine attention to the wide class of hypotheses constraining θ_0 to lie in the null space of some specified vector function. For instance, the discussion of this section will involve two nested hypotheses

$$\omega_1 = \{\theta : \mathbf{h}_1(\theta) = \mathbf{0}\} \supset \omega_2 = \{\theta : \mathbf{h}_2(\theta) = \mathbf{0}\},$$

where $\mathbf{h}_1 = [h_1, h_2, \ldots, h_{r_1}]'$ and $\mathbf{h}_2 = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_{21} \end{bmatrix} = [h_1, h_2, \ldots, h_{r_2}]'$ with $r_2 > r_1$. Three matrices of derivatives are defined:

$\mathbf{H}_{1\theta}$ of order $s \times r_1$ with (i,j)th component $\partial h_j(\theta)/\partial \theta_i$
$(i = 1, 2, \ldots, s; \quad j = 1, 2, \ldots, r_1)$;

$\mathbf{H}_{2\theta}$ of order $s \times r_2$ with (i,j)th component $\partial h_j(\theta)/\partial \theta_i$
$(i = 1, 2, \ldots, s; \quad j = 1, 2, \ldots, r_2)$;

$\mathbf{H}_{21\theta}$ of order $s \times (r_2 - r_1)$ with (i,j)th component $\partial h_j(\theta)/\partial \theta_i$
$(i = 1, 2, \ldots, s; \quad j = r_1 + 1, \ldots, r_2)$,

so that $\mathbf{H}_{2\theta} = [\mathbf{H}_{1\theta} \mathbf{H}_{21\theta}]$, and the matrices are assumed to be of ranks r_1, r_2 and $r_2 - r_1$ respectively. To simplify the notation we use $[\cdot]_\theta$ to denote that the matrix expression in the square brackets is evaluated at θ; thus $[\mathbf{H}_1' \mathbf{B}^{-1} \mathbf{H}_1]_\theta^{-1}$ denotes $[\mathbf{H}_{1\theta}' \mathbf{B}_\theta^{-1} \mathbf{H}_{1\theta}]^{-1}$. We can then introduce three real-valued functions $\lambda_{10}, \lambda_{20}, \lambda_{21}$, each defined on ω_0, by

$$\lambda_{10}(\theta) = n[\mathbf{h}_1'(\mathbf{H}_1' \mathbf{B}^{-1} \mathbf{H}_1)^{-1} \mathbf{h}_1]_\theta,$$

$$\lambda_{20}(\theta) = n[\mathbf{h}_2'(\mathbf{H}_2' \mathbf{B}^{-1} \mathbf{H}_2)^{-1} \mathbf{h}_2]_\theta,$$

$$\lambda_{21}(\theta) = n[\mathbf{h}_{21}'\{\mathbf{H}_{21}' \mathbf{B}^{-1} \mathbf{H}_{21} - \mathbf{H}_{21}' \mathbf{B}^{-1} \mathbf{H}_1 (\mathbf{H}_1' \mathbf{B}^{-1} \mathbf{H}_1)^{-1} \mathbf{H}_1' \mathbf{B}^{-1} \mathbf{H}_{21}\}^{-1} \mathbf{h}_{21}]_\theta.$$

3.4. *Chi-squared Points*

The upper α point of the $\chi^2[r]$ distribution is denoted by $c_{r,\alpha}$ so that

$$\Pr\{\chi^2[r] > c_{r,\alpha}\} = \alpha.$$

Also we use $\chi^2[r, \lambda]$ to denote a non-central chi-squared random variable with r degrees of freedom and non-centrality parameter λ, that is a random variable with moment generating function $(1-2t)^{-\frac{1}{2}r} \exp\{\lambda t/(1-2t)\}$ $(|t| < \frac{1}{2})$.

The details of the three large sample equivalent tests of size α of the hypothesis ω_1 against the complete alternative $\omega_0 - \omega_1$ are as follows.

3.5. *Generalized Likelihood Ratio Test*

This test uses a test statistic $T_{10} = 2\log\{L(\cdot, \theta^0)/L(\cdot, \theta^1)\}$ and critical region $\{x : T_{10}(x) > c_{r_1,\alpha}\}$ of size α. For its application we have to compute *both* θ^0 and θ^1, and to do so generally requires the application of at least one iterative procedure involving the evaluation of some B_θ^{-1}.

3.6. *Lagrange-multiplier Test*

Here the test statistic is $V_{10} = (1/n)\mathbf{D}' \log L(\cdot, \theta^1) \mathbf{B}_1^{-1} \mathbf{D} \log L(\cdot, \theta^1)$ and the critical region $\{x : V_{10}(x) > c_{r_1,\alpha}\}$. The main quantities to be computed are θ^1, which generally requires an iterative determination, and $B_{\theta^1}^{-1}$.

3.7. *Wald Test*

The Wald test statistic is defined by $W_{10} = \lambda_{10}\{\theta^0(\cdot)\}$ and the corresponding critical region is $\{x : W_{10}(x) > c_{r_1,\alpha}\}$. The computation of θ^0 and $B_{\theta^0}^{-1}$—is often simple in applications and provided r_1 is not too large the evaluation of $[\mathbf{H}_1' \mathbf{B}^{-1} \mathbf{H}_1]_{\theta^0}^{-1}$ may be easy.

3.8. *Discussion*

The three test statistics are distributed as $\chi^2[r_1]$ when $\theta_0 \in \omega_1$, as is evident by the form of the critical region, and as $\chi^2[r_1, \lambda_{10}(\theta_0)]$ when $\theta_0 \in \omega_0 - \omega_1$. The power function β of the tests is therefore given by $\beta(\theta_0) = \Pr\{\chi^2[r_1, \lambda_{10}(\theta_0)] > c_{r_1,\alpha}\}$.

The advantages of the last two tests over the generalized likelihood ratio test from the computational point of view, and also the relative merits of these two tests, have recently been the subject of a review by Aitchison and Silvey (1960). When we attempt to apply the Lagrange-multiplier and the Wald techniques to derive tests of ω_2 against $\omega_1 - \omega_2$ within ω_0 we can encounter difficulties not usually met in the application of the procedures to straightforward unrestricted tests. Both techniques require us to derive and invert the information matrix associated with the adopted model ω_1. This involves the recognition that the specialization of ω_0 to ω_1 effectively reduces the number of independent parameters from s to $s - r_1$ and so a reparametrization of the model in terms of such a set of $s - r_1$ parameters is necessary before we can proceed. This in itself may be a formidable task. Even when it is possible, as it is in all our illustrations (chosen for their simplicity rather than their reality), further snags lie ahead. For not only can the components of the information matrix associated with ω_1 be awkward to derive and compute but this information matrix will seldom possess as simple a form as that associated with the corresponding unrestricted test. In Example 2, for instance, probably the easiest way of reparametrizing is to introduce the set of $k+1$ new independent parameters $\phi_1, \phi_2, \ldots, \phi_{k+1}$ by $\theta_{2i-1} = \phi_i$, $\theta_{2i} = \phi_i \phi_{k+1}$ $(i = 1, 2, \ldots, k)$. The information matrix associated with ω_1, expressed in terms of ϕ, is no longer diagonal as is the information matrix associated with ω_0 but is of bordered diagonal form. The fact that the order of the matrix has been reduced from $2k$ to $k+1$ does not compensate for this complication at the inversion stage. This extra complexity of information matrices associated with restricted tests adds considerably to the computational effort required in the direct application of the two techniques.

For the Wald technique there is at first sight a compensating feature in that the new matrix of form $\mathbf{H}'\mathbf{B}^{-1}\mathbf{H}$ is of smaller order, and so may pose an easier inversion problem, than the corresponding matrix for the unrestricted case. This is almost certainly countered by the need to obtain θ^1, whose evaluation, as compared with that of θ^0, generally requires computations of greater difficulty; very likely an iterative procedure will be required. Even in the extremely simple case of Example 1 we have $\theta^0 = (\bar{x}, s)$, where $n\bar{x} = \Sigma x_i$ and $ns^2 = \Sigma(x_i - \bar{x})^2$, whereas $\theta^1 = (a, a)$, where a is the positive root of the quadratic equation $\alpha^2 + \bar{x}\alpha - (\bar{x}^2 + s^2) = 0$. As an alternative, which avoids reparametrization, to the above Wald procedure we can show that the Wald statistic W_{21} for testing ω_2 against $\omega_1 - \omega_2$ within ω_0 can be expressed in the form $W_{21} = \lambda_{21}\{\theta^1(\cdot)\}$, but again this does not present easy computations.

These difficulties, the necessary reparametrization of the model, the more intricate structure of the information matrix and the necessity to employ more involved computational procedures to obtain maximum likelihood estimates, indicate that it may be fruitful to re-examine the techniques. This is undertaken in the next section, where we aim to provide a whole variety of equivalent tests; which is used is then a question of which is computationally most convenient for the special features of the particular problem under discussion.

3.9. *Note on Singular Information Matrices*

When $\mathbf{B}_\theta$ is not non-singular for all $\theta \in \omega_0$ the question of the identifiability of parameters arises. Usually, then, if the first r_0 constraints $h_1(\theta) = h_2(\theta) = \ldots h_{r_0}(\theta) = 0$

of $\mathbf{h}_1(\theta) = \mathbf{0}$ are just sufficient to identify the parameters and $\mathbf{H}_{0\theta}$ is the corresponding $s \times r_0$ matrix of first derviatives associated with these constraints, the matrix $[\mathbf{B}+\mathbf{H}_0\mathbf{H}_0']_\theta$ is non-singular for all $\theta \in \omega_0$, and all that is necessary in the above tests of ω_1 against $\omega_0 - \omega_1$ is to replace $\mathbf{B}$ by $\mathbf{B}+\mathbf{H}_0\mathbf{H}_0'$ and the number of degrees of freedom r_1 by $r_1 - r_0$. These changes in no way affect the discussion of difficulties above; indeed they will usually make the application of the techniques, as they at present stand, more involved.

4. Basic Results on Restricted Tests

Theorem 1 below, by establishing the large-sample equivalence of a number of critical regions, provides the basis for a choice of computational routines in carrying out a restricted test of ω_2 against $\omega_1 - \omega_2$ within ω_0.

Theorem 1. Let T_{ji}, V_{ji}, W_{ji} $(ji = 10, 20, 21)$ be respectively the generalized likelihood-ratio, the Lagrange-multiplier and the Wald statistics for testing ω_j against $\omega_i - \omega_j$. If A_{ji} and B_{ji} denote any of T_{ji}, V_{ji}, W_{ji}, then the twelve tests, of which three have critical regions

$$\{x : A_{21}(x) > c_{r_2-r_1,\alpha}\}$$

and nine have critical regions

$$\{x : A_{20}(x) - B_{10}(x) > c_{r_2-r_1,\alpha}\},$$

are large-sample equivalent restricted tests (of size α) of ω_2 against $\omega_1 - \omega_2$ within ω_0. The power function β of these tests is given by

$$\beta(\theta_0) = \Pr\{\chi^2[r_2 - r_1, \lambda_{20}(\theta_0)] > c_{r_2-r_1,\alpha}\}.$$

Proof. The equivalence of the three tests based on T_{21}, V_{21}, W_{21} is an established result of unrestricted theory.

Clearly from the definition of T_{ji} we have $T_{21} = T_{20} - T_{10}$ and thus to establish the equivalence property all that we require to prove is that $A_{20} - B_{10}$ and $T_{20} - T_{10}$ have the same asymptotic distribution when $\theta_0 \in \omega_1$. To establish this, methods of proof similar to those of Silvey (1959) give us, for each $\theta_0 \in \omega_1$,

$$\operatorname{plim}(V_{ji} - T_{ji}) = 0,$$

$$\operatorname{plim}(W_{ji} - T_{ji}) = 0,$$

from which we have

$$\operatorname{plim}\{(A_{20} - B_{10}) - (T_{20} - T_{10})\} = 0,$$

and the required equivalence of the asymptotic distributions of $A_{20} - B_{10}$ and $T_{20} - T_{10}$ follows immediately, for example, by the results of Mann and Wald (1943).

From unrestricted theory, we know that

$$\beta(\theta_0) = \Pr\{\chi^2[r_2 - r_1, \lambda_{21}(\theta_0)] > c_{r_2-r_1,\alpha}\} \quad (\theta_0 \in \omega_1 - \omega_2),$$

and so the result of the theorem follows if we can show that $\lambda_{21}(\theta_0) = \lambda_{20}(\theta_0)$ for $\theta_0 \in \omega_1 - \omega_2$. Now for $\theta_0 \in \omega_1 - \omega_2$ we have

$$\begin{aligned}
\lambda_{20}(\theta_0) &= [\mathbf{h}_2'(\mathbf{H}_2'\mathbf{B}^{-1}\mathbf{H}_2)^{-1}\mathbf{h}_2]_{\theta_0} \\
&= [\mathbf{0}\ \mathbf{h}_{21}'(\theta_0)] \begin{bmatrix} \mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_1 & \mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_{21} \\ \mathbf{H}_{21}'\mathbf{B}^{-1}\mathbf{H}_1 & \mathbf{H}_{21}'\mathbf{B}^{-1}\mathbf{H}_{21} \end{bmatrix}_{\theta_0}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{h}_{21}(\theta_0) \end{bmatrix} \\
&= \mathbf{h}_{21}'(\theta_0)[\mathbf{H}_{21}'\mathbf{B}^{-1}\mathbf{H}_{21} - \mathbf{H}_{21}'\mathbf{B}^{-1}\mathbf{H}_1(\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_1)^{-1}\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_{21}]_{\theta_0}^{-1}\mathbf{h}_{21}(\theta_0) \\
&= \lambda_{21}(\theta_0).
\end{aligned}$$

In any application we naturally choose the test statistic which can be most readily computed. Frequently this turns out to be $W_{20}-W_{10}$, for often θ^0 and $B_{\theta^0}^{-1}$ are easy to evaluate and the remaining calculations are not too heavy. In Example 1, for instance, $W_{20}(x)-W_{10}(x) = \frac{1}{3}n(\bar{x}+2s-3)^2/s^2$ and this is more easily computed than any of the other equivalent statistics. Again, in Example 2, θ^0 and $B_{\theta^0}^{-1}$ are easily obtained and although the use of $W_{20}-W_{10}$ requires further the inversion of matrices of order k and $k+1$ this statistic, if k is not too large, is probably easier to calculate than any of the others which require iterative evaluation of maximum likelihood estimates. In Example 4, however, for the restricted test of ω_3 against $\omega_1-\omega_3$ the statistic $V_{30}-W_{10}$ would seem a sensible one, since $\theta^3 = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ requires no determination and V_{30}, being in this case the usual chi-squared goodness-of-fit test statistic of ω_3 against $\omega_0-\omega_3$, is easily obtained; also the calculation of W_{10}, avoiding reparametrization, is probably the easiest way of handling the adopted model ω_1.

To provide a numerical comparison of the twelve test statistics we have evaluated them for the cases:

(i) Example 1 with $n = 100$, $\bar{x} = 1{\cdot}2$, $s = 1{\cdot}1$;

(ii) Example 1 with $n = 100$, $\bar{x} = 1{\cdot}15$, $s = 1{\cdot}05$;

(iii) Example 3 with $p = 2$ and observed mean counts over 100 replicates from the three sources of 1·7, 4·8, 8·3. In hypothesis ω_2 we suppose $d = 3$.

TABLE 1

Values of equivalent test statistics

Test statistic	(i)	(ii)	(iii)
$T_{21} = T_{20}-T_{10}$	5·42	2·17	3·14
$T_{20}-V_{10}$	5·44	2·24	3·15
$T_{20}-W_{10}$	5·39	2·14	3·14
V_{21}	6·75	2·52	3·19
$V_{20}-T_{10}$	6·60	2·29	3·19
$V_{20}-V_{10}$	6·63	2·36	3·20
$V_{20}-W_{10}$	6·57	2·26	3·20
W_{21}	4·24	1·84	3·11
$W_{20}-T_{10}$	4·44	1·92	3·05
$W_{20}-V_{10}$	4·47	1·99	3·06
$W_{20}-W_{10}$	4·41	1·89	3·05

There is little variation in the values of the test statistics in case (iii). In cases (i) and (ii) while there is agreement between statistics based on the same leading statistic the values may vary from group to group. Whether this will often in practice result in the tests yielding different decisions is still an open question which is not, however, peculiar to restricted tests but to the whole field of large-sample equivalent tests. In the absence of any definite answer to this question the only sensible advice to give to users of large-sample equivalent tests is to use whichever is computationally most convenient and to treat with caution borderline decisions.

5. APPLICATIONS

The wide variety of equivalent test statistics provided by the theorem of the previous section is the necessary apparatus for the handling of the applications of the restricted tests as outlined in section 2.

5.1. *Analysis of Discrepancy for Nested Hypotheses*

For the analysis of a sequence of nested hypotheses $\omega_0 \supset \omega_1 \supset \omega_2 \supset \ldots \supset \omega_k$, where $\omega_i = \{\theta : \mathbf{h}_i(\theta) = \mathbf{0}\}$ with $\mathbf{h}_i = [h_1, h_2, \ldots, h_{r_i}]'$ and $r_j > r_i$ for $j > i$, Wald test statistics of the type $W_{j0} - W_{i0}$ $(j > i)$ for the restricted test of ω_j against $\omega_i - \omega_j$ are admirably suited. For, by using these statistics we base all the tests on θ^0 and the inverse $B_{\theta^0}^{-1}$ of the information matrix associated with ω_0, and although we require a matrix inversion in the computation of each $W_{i0}(x)$ the work involved is generally slight compared with

TABLE 2

Analysis of discrepancy layout

		Hypothesis ω_i of comparison		
		$0(r_0)$†	$1(r_1)$	$2(r_2)$ ...
Hypothesis ω_j under test	$1(r_1)$	$W_{10}(x)$		
	$2(r_2)$	$W_{20}(x)$	$W_{20}(x) - W_{10}(x)$	
	$3(r_3)$	$W_{30}(x)$	$W_{30}(x) - W_{10}(x)$	$W_{30}(x) - W_{20}(x)$
	⋮			

† The number of restrictions in each hypothesis is shown in brackets, and $r_0 = 0$ if $\mathbf{B}_\theta$ is non-singular for all $\theta \in \omega_0$; otherwise r_0 is the number of restraints required for identifiability.

the otherwise necessary reparametrization and the calculation of other maximum likelihood estimates and information matrices. The results of an analysis may be displayed in an analysis of discrepancy table, the jth row of the table presenting the values $W_{j0}(x)$ and $W_{j0}(x) - W_{i0}(x)$ $(i = 1, 2, \ldots, j-1)$, which we may regard as measures, supplied by the data, of the discrepancies between the hypothesis ω_j under test and the hypothesis ω_i of comparison. Only if all the values in a row are non-significant (the assessment being made at the appropriate number of degrees of freedom easily noted as $r_j - r_i$ by reference to the margins of the table) do we accept hypothesis ω_j and move to the next row. Only one computation, namely that of $W_{j0}(x)$, is necessary for each row of the table, the other values being obtained by simple subtractions. We remind the reader at this stage of our earlier remarks about the multiplicity of significance tests involved in this analysis.

TABLE 3

Analysis of discrepancy table

		Hypothesis of comparison ω_i		
		0(0)	1(1)	2(2)
Hypothesis ω_j under test	1(1)	0·55		
	2(2)	3·60	3·05	
	3(3)	7·21	6·66	3·61

Illustration. The method is amply illustrated by constructing the analysis of discrepancy table for the nested hypotheses $\omega_0 \supset \omega_1 \supset \omega_2 \supset \omega_3$ of Example 3, where $a = 2$, $d = 3$, $p = 2$. We suppose that, as in our previous illustration, the mean counts from the three sources over 100 replicates are $\bar{x}_1 = 1{\cdot}7$, $\bar{x}_2 = 4{\cdot}8$, $\bar{x}_3 = 8{\cdot}3$. The computations are very simple in this case since $\mathbf{B}_\theta^{-1} = \text{diag}\{\theta_1, \theta_2, \theta_3\}$ and $\theta^0 = \{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$. It is perhaps worth noting that the evaluation of W_{30} can be greatly simplified by rewriting $\omega_3 = \{\theta : \theta_1 - 2 = \theta_2 - 5 = \theta_3 - 8 = 0\}$, a device which is often rewarding.

In the table the only significant value at the 5 per cent. level is $W_{30}(x)-W_{10}(x) = 6{\cdot}66$ which is greater than $c_{3-1,0{\cdot}05} = 5{\cdot}84$. Thus we would be prepared to go no further than to accept ω_2 on this data. The lengthier process of carrying out the analysis using statistics other than this Wald type leads to exactly the same conclusions.

5.2. *Increasing Power by Restricting Alternatives*

The power at $\theta_0 \in \omega_1-\omega_2$ of a restricted test of ω_2 against $\omega_1-\omega_2$ within ω_0 is

$$\Pr\{\chi^2[r_2-r_1, \lambda_{20}(\theta_0)] > c_{r_2-r_1,\alpha}\},$$

as compared with the power of the corresponding unrestricted test of ω_2 against $\omega_0-\omega_2$ which is given by

$$\Pr\{\chi^2[r_2, \lambda_{20}(\theta_0)] > c_{r_2,\alpha}\}.$$

The two expressions differ only in the degrees of freedom involved, namely r_2-r_1 and r_2. A similar result is established by a geometrical argument for their special multinomial case by Fix, Hodges and Lehmann (1959). Their conclusion, reached from graphical considerations of the power functions, also holds here, namely that for fixed $\theta_0 \in \omega_1-\omega_2$ the restricted test power is greater than the unrestricted test power and the difference increases as r_1 increases. Thus when r_1 is small the increase in power may not be great whereas when r_1 is larger the increase may be more appreciable. This point may be illustrated by our examples.

Example 1. For $n = 100$ and $\theta_0 \in \omega_1-\omega_2$ with $\theta_1 = \theta_2 = 0{\cdot}9$ we have $\lambda_{20}(\theta_0) = \lambda_{21}(\theta_0) = 3{\cdot}70$ and the restricted test of ω_2 against $\omega_1-\omega_2$ within ω_0 has power 0·48 compared with the unrestricted test power of 0·38; for $\theta_0 \in \omega_1-\omega_2$ with $\theta_1 = \theta_2 = 1{\cdot}2$ we have $\lambda_{20}(\theta_0) = \lambda_{21}(\theta_0) = 8{\cdot}33$ and the corresponding powers are 0·83 and 0·74. Here $r_1 = 1$ and the increases in power are not very marked. To illustrate the fact that the restricted test may have the smaller power for alternatives in $\omega_0-\omega_1$ we consider the case $n = 100$, $\theta_1 = 1{\cdot}1$, $\theta_2 = 1{\cdot}2$, when the powers for the restricted and unrestricted tests are respectively 0·29 and 0·62.

Example 3. For $p = 8$ the powers of the restricted and unrestricted tests of size 0·05 of ω_2 are respectively 0·81 and 0·53 when $\theta_0 \in \omega_1-\omega_2$ and $\lambda_{20}(\theta_0) = 4$. Here $r_1 = 7$ and the increase in power is more noticeable.

5.3. *Easing the Evaluation of a Test Statistic*

If the information matrix and maximum likelihood estimates associated with ω_1 are difficult to handle it may well be a legitimate computational device in carrying out an unrestricted test of ω_2 against all alternatives within the basic model ω_1 to use a restricted test statistic such as $W_{20}-W_{10}$, where ω_0 is a model such that $\omega_0 \supset \omega_1$. For instance, in Example 1, if we know that X is $N(\theta, \theta^2)$, or that the basic model is ω_1, and wish to test the hypothesis ω_2 that X is $N(1, 1)$ the introduction of ω_0 and the use of $W_{20}(x)-W_{10}(x) = \frac{1}{3}n(\bar{x}+2s-3)^2/s^2$ leads to simpler computations and an equivalent test.

5.4. *The Multinomial Case*

When X is a k-class multinomial random variable with class probabilities $\theta_i/\Sigma\theta_j$ $(i = 1, 2, \ldots, k)$ and $\omega_0 = \{\theta : \text{all } \theta_i \geqslant 0\}$ with the identifiability condition $\Sigma\theta_j = 1$, the Lagrange-multiplier statistic V_{10} for testing ω_1 against $\omega_0-\omega_1$ is then just the usual chi-squared goodness-of-fit test statistic based on the familiar $\Sigma\{(\text{observed}-\text{expected under } \omega_1)^2/(\text{expected under } \omega_1)\}$; see, for example, Aitchison

and Silvey (1960). If in applying a restricted test of ω_2 against $\omega_1-\omega_2$ within ω_0 we use the test statistic $V_{20}-V_{10}$ we then have essentially the restricted chi-squared tests of Fix, Hodges and Lehmann (1959). The multinomial theory thus forms a special case of our more general theory. Moreover, the alternative test statistics provided by Theorem 1 may give easier means of computation than the Fix–Hodges–Lehmann statistics. In testing ω_2 against $\omega_1-\omega_2$ in Example 4 we have an instance of this, for $W_{20}(x)-W_{10}(x)$, although involving the inversion of matrices of order k and $2k$, requires less computational skill than the iterative procedures of finding θ^1 and θ^2 necessary for the $V_{20}-V_{10}$ approach.

6. Separable Hypotheses and Partitioning of Test Statistics

6.1. *Criteria for Separability*

Our remaining task is to demonstrate the application of restricted test theory to the related problems of recognizing separable hypotheses and partitioning test statistics. We state our two results—Theorem 2 for non-singular, and Theorem 3 for singular, information matrices associated with the basic model ω_0—as sufficient conditions for two hypotheses ω_1 and ω_2 to be separable with respect to the equivalent large sample tests of Theorem 1. These give easily applied criteria for separability and at the same time provide the necessary basis for partitioning a test statistic. The theory is most easily developed in terms of Wald statistics, but equivalent statistics can be substituted in the applications whenever convenient.

6.2. *Case of a Non-singular Information Matrix*

We suppose that $\omega_1=\{\theta\in\omega_0 : \mathbf{h}_1(\theta)=\mathbf{0}\}$ and $\omega_2=\{\theta\in\omega_0 : \mathbf{h}_2(\theta)=\mathbf{0}\}$, where $\mathbf{h}_1=[h_{11}h_{12}\dots h_{1r_1}]'$ and $\mathbf{h}_2=[h_{21}h_{22}\dots h_{2r_2}]'$, are no longer nested and let $\mathbf{H}_{1\theta}$ and $\mathbf{H}_{2\theta}$, the $s\times r_1$ and $s\times r_2$ matrices of first order partial derviatives of $\mathbf{h}_1$ and $\mathbf{h}_2$ with respect to θ, be such that $[\mathbf{H}_1\,\mathbf{H}_2]_\theta$ is of rank r_1+r_2 for all $\theta\in\omega_0$. We then have the following theorem for the case where the information matrix based on ω_0 is non-singular.

Theorem 2. If $\mathbf{B}_\theta$ is non-singular for all $\theta\in\omega_0$ then a sufficient condition for ω_1 and ω_2 to be separable with respect to the equivalent large sample tests of Theorem 1 is that $[\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_2]_\theta=\mathbf{0}$ for all $\theta\in\omega_1\cap\omega_2$.

Proof. Let $\omega_3=\omega_1\cap\omega_2$, assumed to be non-empty. To establish the theorem we have to show, for example, that the large sample critical region for testing ω_3 against $\omega_1-\omega_3$ is the same as that for testing ω_2 against $\omega_0-\omega_2$. (The equivalence of the critical regions for testing ω_3 against $\omega_2-\omega_3$ and ω_1 against $\omega_0-\omega_1$ is proved in a similar fashion.) This will be so if we can show that $\operatorname{plim}(W_{31}-W_{20})=0$ if $\theta_0\in\omega_1\cap\omega_2$. Now

$$W_{30}=[\mathbf{h}_1'\mathbf{h}_2']_{\theta^*}\begin{bmatrix}\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_1 & \mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_2\\ \mathbf{H}_2'\mathbf{B}^{-1}\mathbf{H}_1 & \mathbf{H}_2'\mathbf{B}^{-1}\mathbf{H}_2\end{bmatrix}_{\theta^*}^{-1}\begin{bmatrix}\mathbf{h}_1\\ \mathbf{h}_2\end{bmatrix}_\theta$$

and

$$\begin{aligned}\operatorname{plim} W_{30}&=\operatorname{plim}[\mathbf{h}_1'\mathbf{h}_2']_{\theta^*}\begin{bmatrix}[\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_1]_{\theta^*} & [\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_2]_{\theta_*}\\ [\mathbf{H}_2'\mathbf{B}^{-1}\mathbf{H}_1]_{\theta_*} & [\mathbf{H}_2'\mathbf{B}^{-1}\mathbf{H}_2]_{\theta^*}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{h}_1\\ \mathbf{h}_2\end{bmatrix}_{\theta^*}\\ &=\operatorname{plim}[\mathbf{h}_1'\mathbf{h}_2']_{\theta^*}\begin{bmatrix}\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{H}_2'\mathbf{B}^{-1}\mathbf{H}_2\end{bmatrix}_{\theta^*}^{-1}\begin{bmatrix}\mathbf{h}_1\\ \mathbf{h}_2\end{bmatrix}_{\theta^*}\\ &=\operatorname{plim}(W_{10}+W_{20}).\end{aligned}$$

Thus, for $\theta_0 \in \omega_1 \cap \omega_2$, $\quad \operatorname{plim} W_{20} = \operatorname{plim}(W_{30} - W_{10})$

$$= \operatorname{plim} W_{31}$$

by Theorem 1. Hence the theorem is established.

The practical use of the result requires some explanation. When faced with two hypotheses ω_1 and ω_2 as above, the first step, in examining the two nested sequences $(\omega_0, \omega_1, \omega_1 \cap \omega_2)$ and $(\omega_0, \omega_2, \omega_1 \cap \omega_2)$ as indicated in section 2, is to test ω_1 against $\omega_0 - \omega_1$, and ω_2 against $\omega_0 - \omega_2$. Only if these tests accept ω_1 and ω_2 do we inquire whether we should then also accept $\omega_3 = \omega_1 \cap \omega_2$. We would then want to test ω_3 against each of $\omega_1 - \omega_3$, $\omega_2 - \omega_3$ and $\omega_0 - \omega_3$. The above theorem shows that when ω_1 and ω_2 are separable we can omit the tests against $\omega_1 - \omega_3$ and $\omega_2 - \omega_3$ and concentrate on ω_3 against $\omega_0 - \omega_3$. Indeed there are further implications. For acceptance of ω_1 and ω_2 separately means that $W_{10}(x) < c_{r_1,\alpha}$ and $W_{20}(x) < c_{r_2,\alpha}$ and so $W_{10}(x) + W_{20}(x) < c_{r_1,\alpha} + c_{r_2,\alpha}$. Also under the separability conditions of Theorem 2 the critical region

$$\{x : W_{10}(x) + W_{20}(x) > c_{r_1+r_2,\alpha}\}$$

may be used for the test of ω_3 against $\omega_0 - \omega_3$. Now since $c_{r_1,\alpha} + c_{r_2,\alpha}$ is never greatly in excess of $c_{r_1+r_2,\alpha}$ the ready acceptance of ω_1 and ω_2 will usually result in the acceptance of ω_3. Thus the final test is little more than a safeguard against the too eager acceptance of ω_3 after the near rejection of ω_1 and ω_2.

This point can be illustrated by a familiar situation in regression analysis. Suppose that we have n observations on $(X_1, X_2, \ldots, X_k)$, where the components are independent and X_i is distributed as $N(\theta_1 a_i + \theta_2 b_i, \sigma^2)$ $(i = 1, 2, \ldots, k)$, the a_i and b_i being known constants. If $\omega_1 = \{\theta : \theta_1 = 0\}$ and $\omega_2 = \{\theta : \theta_2 = 0\}$ then the separability condition is easily seen to be $\Sigma a_i b_i = 0$, the familiar orthogonality condition. In such circumstances acceptance of ω_1 and ω_2 separately will usually result in acceptance of $\omega_1 \cap \omega_2 = \{\theta : \theta_1 = \theta_2 = 0\}$. On the other hand, if $\Sigma a_i b_i \neq 0$ the interpretation becomes more difficult, for acceptance of ω_1 and ω_2 separately may not lead to the acceptance of $\omega_1 \cap \omega_2$; both the controllable variables a and b may separately give a reasonable fit to the dependent variable x because of their own interdependence.

As in the illustration of the preceding paragraph our result, applied to the general linear hypothesis model, leads to the usual orthogonality conditions. Our main concern is again in non-standard situations of which we give the following illustration.

Illustration. In Example 5 we have

$$\mathbf{B}_\theta^{-1} = \frac{1}{p}\operatorname{diag}\{\theta_1^2, \theta_2^2, \theta_3^2\},$$

$$\mathbf{H}_{1\theta}' = p^3[\theta_2\theta_3 \quad \theta_3\theta_1 \quad \theta_1\theta_2],$$

$$\mathbf{H}_{2\theta}' = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix},$$

and so $[\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_2]_\theta = p^2\theta_1\theta_2\theta_3[\theta_1 - \theta_2 \quad \theta_1 - \theta_3] = \mathbf{0}$ when $\theta \in \omega_1 \cap \omega_2$. Hence ω_1 and ω_2 are separable. Also $p\bar{\theta}^0 = (\bar{x}_1, \bar{x}_2, \bar{x}_3)$, where $\bar{x}_1, \bar{x}_2, \bar{x}_3$ are the means of the three linear measurements of the n cuboids, and it easily follows that

$$W_{10}(x) = np(\bar{x}_1\bar{x}_2\bar{x}_3 - a^3)^2/(\bar{x}_1\bar{x}_2\bar{x}_3)^2,$$

$$W_{20}(x) = np\left\{\left(\frac{1}{\bar{x}_1} - \frac{1}{\bar{x}_2}\right)^2 + \left(\frac{1}{\bar{x}_1} - \frac{1}{\bar{x}_3}\right)^2 + \left(\frac{1}{\bar{x}_2} - \frac{1}{\bar{x}_3}\right)^2\right\} \Big/ \left(\frac{1}{\bar{x}_1^2} + \frac{1}{\bar{x}_2^2} + \frac{1}{\bar{x}_3^2}\right)$$

and

$$W_{30}(x) = np\left\{\left(1-\frac{a}{\bar{x}_1}\right)^2+\left(1-\frac{a}{\bar{x}_2}\right)^2+\left(1-\frac{a}{\bar{x}_3}\right)^2\right\}.$$

We display in Table 4 the results of the three tests of size 0·05 for the cases
(i) $n = 100$, $p = 3$, $a = 1$; $\bar{x}_1 = 1{\cdot}10$, $\bar{x}_2 = 1{\cdot}05$, $\bar{x}_3 = 0{\cdot}95$;
(ii) $n = 100$, $p = 3$, $a = 1$; $\bar{x}_1 = 1{\cdot}14$, $\bar{x}_2 = 1{\cdot}12$, $\bar{x}_3 = 0{\cdot}96$.

TABLE 4

Values of test statistics

		Case (i)	*Case* (ii)	*Corresponding values of* $c_{r,\alpha}$
Test statistic	W_{10} $(r_1 = 1)$	0·79	3·40	3·84
	W_{20} $(r_2 = 2)$	3·43	5·59	5·99
	W_{30} $(r_1+r_2 = 3)$	3·99	8·30	7·81

Case (i) shows the customary acceptance of $\omega_1 \cap \omega_2$ after the ready acceptance of ω_1 and ω_2 separately; case (ii) shows how near rejection of ω_1 and ω_2 can lead to the rejection of $\omega_1 \cap \omega_2$.

6.3. *Case of a Singular Information Matrix*

Suppose now that $\mathbf{B}_\theta$ can be singular and that the necessary r_0 constraints $\mathbf{h}_0(\theta) = \mathbf{0}$ for identifiability of the parameters have been imposed, with as before the $s \times r_0$ matrix $\mathbf{H}_0$ of first order partial derivatives of the components of $\mathbf{h}_0$ such that $[\mathbf{B}+\mathbf{H}_0\mathbf{H}_0']_\theta$ is non-singular for all $\theta \in \omega_0$. The hypotheses ω_1 and ω_2 are as for the non-singular case. We then have the following criterion for separability of ω_1 and ω_2.

Theorem 3. In the singular case ω_1 and ω_2 are separable with respect to the equivalent large sample tests of Theorem 1 if, for all $\theta \in \omega_1 \cap \omega_2$,

$$[\mathbf{H}_1'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_2]_\theta = \mathbf{0}$$

and either $[\mathbf{H}_0'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_1]_\theta = \mathbf{0}$ or $[\mathbf{H}_0'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_2]_\theta = \mathbf{0}$.

Proof. The proof follows a pattern similar to that of Theorem 2, and we omit the details. The sufficient conditions are introduced at the same stage (together with the fact that $\mathbf{h}_0(\theta^0) = \mathbf{0}$) in the establishing of the basic result that $\operatorname{plim} W_{30} = \operatorname{plim}(W_{10}+W_{20})$ if $\theta_0 \in \omega_1 \cap \omega_2$.

Again through the equivalence of the Wald and the Lagrange-multiplier statistics the result covers the existing theory of the multinomial case; for instance, in a 2×2 contingency table the hypothesis ω_1 that the two classifications are independent and the hypothesis ω_2 that the marginal probabilities are specified numbers are separable. The result applies, however, to more than standard cases. For example, in Example 6 it is easy to show that $[\mathbf{H}_0'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_1]_\theta = \mathbf{0}$, $[\mathbf{H}_0'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_2]_\theta = \mathbf{0}$, $[\mathbf{H}_1'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_2]_\theta = \mathbf{0}$ for all $\theta \in \omega_1 \cap \omega_2$ and so the two hypotheses are separable; we could then proceed to investigate the three tests as in the non-singular test and again we would find the same pattern exhibited. It is worth noting here that if we use the identifiability condition in ω_1 and so consider the equivalent hypothesis $\omega_1' = \{\theta \in \omega_0 : \theta_1^{\frac{1}{2}} + \theta_3^{\frac{1}{2}} = 1\}$ we obtain the same results via a different route. Although

$[\mathbf{H}_0'(\mathbf{B}+\mathbf{H}_0\mathbf{H}_0')^{-1}\mathbf{H}_1]_\theta$ is now no longer **0** for all $\theta \in \omega_1 \cap \omega_2$ the remaining two properties are sufficient to establish the separability of ω_1 and ω_2.

It is clear that the notion of separability can easily be extended to more than two hypotheses but neither here nor in dealing with partitioning shall we enter into the details of such extensions.

6.4. *Partitioning a Test Statistic*

The results of our three theorems may now be applied to the problem of partitioning test statistics into component statistics. If in investigating ω_3 against $\omega_0-\omega_3$ we have found $W_{30}(x)$, say, significant and if $\omega_3 = \omega_1 \cap \omega_2$ then we may be interested in asking whether this significance is due to departure from ω_1 or from ω_2 or from both ω_1 and ω_2. It should be emphasized that the attitude of the experimenter is here different from that of the experimenter of the last section. Here the experimenter, having found a significant test statistic, has realized that it may be possible to investigate in more detail the source of this significance by partitioning the test statistic. The interpretation becomes relatively easy if $W_{30}(x)$ can be expressed as the sum of two components, one testing ω_1 against $\omega_0-\omega_1$ and the other testing ω_2 against $\omega_0-\omega_2$. Such is the case if ω_1 and ω_2 are separable. For, by Theorem 1, we know that $W_{30}(x)$ is approximately $W_{10}(x)+W_{31}(x)$, and, by Theorems 2 and 3, we have that $W_{31}(x)$ which tests ω_3 against $\omega_1-\omega_3$ serves, in the case where ω_1 and ω_2 are separable, as a test statistic for examining ω_2 against $\omega_0-\omega_2$, or equivalently can be replaced by $W_{20}(x)$. Thus when ω_1 and ω_2 are separable we do have this possibility of partitioning (at least in an asymptotic sense) $W_{30}(x)$ into components $W_{10}(x)$ and $W_{20}(x)$ for the study of departures from ω_1 and ω_2 separately. Familiar examples of this technique are the linear form partitionings of a chi-squared goodness-of-fit test statistic (see, for example, Cochran (1952)), and the breakdown of a sum of squares into component sums in an analysis of variance.

Illustration. If in Example 6 the numbers falling into the three types are (n_1, n_2, n_3) in n replicates then we have

$$W_{10}(x) = 2n(n_1^{\frac{1}{2}}+n_3^{\frac{1}{2}}-n^{\frac{1}{2}})^2/\{n-(n_1^{\frac{1}{2}}+n_3^{\frac{1}{2}})^2\},$$

$$W_{20}(x) = n(n_1-n_3)^2/\{n(n_1+n_3)-(n_1-n_3)^2\},$$

$$W_{30}(x) = \frac{(n_1-\frac{1}{4}n)^2}{n_1}+\frac{(n_2-\frac{1}{2}n)^2}{n_2}+\frac{(n_3-\frac{1}{4}n)^2}{n_3}.$$

Suppose that we find that ω_3 is unacceptable as it is in the case $n = 100$, $n_1 = 20$, $n_2 = 40$, $n_3 = 40$, when $W_{30}(x) = 9{\cdot}38$ to be compared with $c_{2,0{\cdot}05} = 5{\cdot}99$. Then $W_{10}(x) = 2{\cdot}54$ and $W_{20}(x) = 6{\cdot}67$, each to be compared with $c_{1,0{\cdot}05} = 3{\cdot}84$. Our conclusion then is that the rejection of ω_3 is not due to genetic instability of the population but entirely to the inequality of the dominant and recessive proportions.

ACKNOWLEDGEMENT

I wish to thank the referee for helpful comments on an earlier draft of this paper.

REFERENCES

AITCHISON, J. and SILVEY, S. D. (1958), "Maximum-likelihood estimation of parameters subject to restraints", *Ann. math. Statist.*, **29**, 813–828.

—— —— (1960), "Maximum-likelihood estimation procedures and associated tests of significance", *J. R. statist. Soc.* B, **22**, 154–171.

COCHRAN, W. G. (1952), "The χ^2 test of goodness of fit", *Ann. math. Statist.*, **23**, 315–345.
FIX, E., HODGES, J. L. and LEHMANN, E. L. (1959), "The restricted chi-squared test", in *Probability and Statistics, The Harald Cramér Volume*; editor, U. Grenander. New York: Wiley.
MANN, H. B. and WALD, A. (1943), "On stochastic limit and order relationships", *Ann. math. Statist.*, **14**, 217–226.
RAO, C. R. (1948), "Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation", *Proc. Camb. phil. Soc.*, **44**, 50–57.
SILVEY, S. D. (1959), "The Lagrange-multiplier test", *Ann. math. Statist.*, **30**, 389–407.
WALD, A. (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large", *Trans. Amer. math. Soc.*, **54**, 426–482.

AITCHISON, J. (1964)

Confidence-region tests

Reprinted from *J. R. Statist. Soc.* B26, 462-76

Confidence-region Tests

By J. AITCHISON

University of Liverpool

[Received April 1964]

SUMMARY

To test a hypothesis that a parameter has value θ^* we usually ask whether an observation falls in a critical region of the outcome space. It is well known that, by a suitable choice of confidence region, an equivalent test is to ask whether θ^* lies outside the confidence region. In this paper the concept of such a confidence-region test is extended in directions which are relevant to the study of data-snooping and multiple-hypothesis testing. These problems are investigated for linear models and, in the asymptotic case, for more general parametric models.

1. INTRODUCTION

SUPPOSE that the family of distributions, which form the possible probabilistic descriptions of an experiment, is indexed by a parameter θ, and that ω_0 is the parameter space. It is not known what the true value of the indexing parameter is. For convenience we identify a subset ω_1 of ω_0 with the hypothesis that the true parameter value is a member of ω_1; in particular the hypothesis θ^* states that the true parameter value is θ^*. Let $A(\theta^*)$ be a critical region of size α for testing the hypothesis θ^*, and let $C(x)$ be a confidence region for the parameter, based on the observation or outcome x and having confidence coefficient $1-\alpha$. Then it is well known (Neyman, 1937; Lehmann, 1959, pp. 78–83) that, under favourable circumstances, the following two statements are equivalent:

(i) $x \in A(\theta^*)$, and so hypothesis θ^* is rejected;
(ii) $\theta^* \notin C(x)$.

This equivalence can, in fact, always be arranged by selecting

$$C(x) = \{\theta : x \notin A(\theta)\}.$$

In such circumstances we can carry out a test of the hypothesis θ^* according to the following procedure:

(a) If $\{\theta^*\} \cap C(x) = \emptyset$, then reject θ^*;
(b) If $\{\theta^*\} \cap C(x) \neq \emptyset$, then conclude that there is no evidence to justify the rejection of θ^*.

Expressed in other words, a confidence region for the parameter and critical regions of tests of simple hypotheses are here so related that points outside the confidence region correspond to simple hypotheses which would be rejected under the tests. Such an equivalence may also hold when the hypothesis concerns θ in a simple sense but when there are other nuisance parameters involved. A familiar example is the following. Let the observation be the set $x = (x_1, \ldots, x_n)$ of outcomes of n independent replicates of an experiment described by some $N(\theta, \sigma^2)$ distribution, and let interest be in the hypothesis that $\theta = \theta^*$, so that σ is a nuisance parameter. If $\bar{x}$ and s are

defined by $n\bar{x} = \Sigma x_i$, $(n-1)s^2 = \Sigma(x_i - \bar{x})^2$, we have the usual critical region $A(\theta^*)$ of size α for testing θ^* given by

$$A(\theta^*) = \{x: n^{\frac{1}{2}}|\bar{x} - \theta^*|/s > t(n-1; 1-\tfrac{1}{2}\alpha)\},$$

where $t(d; p)$ denotes the p-probability point of the t-distribution with d degrees of freedom; and the usual confidence interval for θ with coefficient $1-\alpha$ is given by

$$C(x) = \{\theta : \bar{x} - sn^{-\frac{1}{2}}t(n-1; 1-\tfrac{1}{2}\alpha) < \theta < \bar{x} + sn^{-\frac{1}{2}}(n-1; 1-\tfrac{1}{2}\alpha)\}.$$

It is clear that $x \in A(\theta^*)$ if and only if $\theta^* \notin C(x)$, as required for the equivalence of the two procedures.

In this paper we consider some of the consequences of regarding an extension of the confidence-region procedure set out in (a) and (b) as the basic method of testing. Our method of bringing experimental evidence x to bear on a possibly composite hypothesis ω_1 will be to ask whether a confidence region $C(x)$, based on x and selected with the experimental setting in mind, is such that

(a) $\omega_1 \cap C(x) = \emptyset$, in which case ω_1 is rejected,
(b) $\omega_1 \cap C(x) \neq \emptyset$, in which case ω_1 is not rejected.

This approach to testing by way of confidence-region tests has repercussions mainly in two aspects of statistical hypothesis testing, namely in the problem of data-snooping and in situations where more than two hypotheses (a null and an alternative) come under scrutiny. Thus a main consideration of this paper is to show the relevance of the technique to these two statistical problems. The other important consideration is the purely mathematical task of transforming the procedure laid down in (a) and (b) into one which is computationally operable. We shall see that the result of this transformation leads to a useful interpretation in terms of common practice.

The use of confidence-region tests is certainly not new in linear statistical analysis. In the analysis of variance they form the basis of the *S*-method of multiple comparison (Scheffé, 1953; 1959, pp. 68–83) for discovering which estimable functions of a specified linear space of estimable functions are significantly different from zero. Our reappraisal of the *S*-method involves certain new features. Whereas the standard method of application consists of asking whether individual estimable functions are significantly different from zero, it is necessary, and mathematically no more difficult, in a general study of the properties of the tests, to ask whether a given linear subspace of the space of estimable functions contains any functions which differ significantly from zero. The resolution of the computational aspect of this problem leads to an exceedingly simple test procedure. Moreover, the theory allows a comparison to be made between the *S*-method and the still prevalent "fixed-level" test procedure, by providing a measure which can conveniently be interpreted by the fixed-level tester as an "equivalent significance level", at which the *S*-method appears to operate. Besides this extension of the results of Scheffé we also advocate a more extensive use of the method.

Confidence-region tests have also recently (Goodman, 1964) been made the basis of multiple comparison in contingency tables, and H. Stone, in the discussion on Beale (1960), indicates how the *S*-method can be applied to regression models which are non-linear in the parameters. These applications are in fact particular cases of a very general theory of large-sample multiple-hypothesis testing in a parametric model, the large-sample analogue of the *S*-method. The essentials of this general theory are presented in this paper.

2. Some Remarks on Data-snooping and Multiple-hypothesis Testing

2.1. *Testing of Hypotheses*

So that the reader may more readily evaluate the procedures we advocate later, we first comment on the philosophy of hypothesis testing adopted in this paper. We regard hypothesis testing as an attempt to reduce the set ω_0 of uncertainty about the true parameter value by contriving to use the observed data to dismiss from the field of possibility subsets of ω_0, that is to say, hypotheses. This process can usefully be looked on as an interchange of views between an experimenter and a statistician. The experimenter presents the statistician with ω_1 and x and asks whether he can now reasonably limit his attention to the set $\omega_0-\omega_1$. Valuable conclusions are most likely to be drawn when this null hypothesis ω_1—selected prior to experimentation—is not so "large" that it is almost impossible to reject it, and not so "small" that to dismiss it is to leave the field of uncertainty scarcely narrowed. This is essentially the outlook adopted by the Neyman–Pearson school who, by using test procedures which protect the null hypothesis through limiting the liability to commit an error of the first type, ensure that rejection of a null hypothesis is a relatively positive action. An experimenter who is delighted when a hypothesis ω_1, suggested by him, is not so rejected should be cautioned to study the power function of the test. We know that often "acceptance" of ω_1 does not imply dismissal of $\omega_0-\omega_1$, but rather that the experimenter has still grounds for hoping that the true parameter value may be in ω_1. If the experimenter really wants to convince himself that ω_1 is acceptable and use the customary type of statistical test, he would be well advised to treat $\omega_0-\omega_1$ as null hypothesis. Procedures are also logically less soundly based when it is not clear that the experimenter selected ω_1 prior to experimentation, or when an answer to the first question of the experimenter is followed by further questions about other hypotheses. We shall see that confidence-region tests have a contribution to make in this troubled area of statistical activity.

It is a generally accepted tenet of scientific method that it is incorrect to use experimental data either to support or refute a hypothesis suggested by those data. Yet we as statisticians, in a variety of ways and to varying degrees, break this tenet unwittingly in much of our routine work, especially when more than one hypothesis is concerned. We mention here a few situations where tests cannot be interpreted in the way that the experimenter is often led to believe that they can.

2.2. *Data-snooping*

Most statisticians, who have been involved to any extent in advisory or consultative work, are familiar with the type of client who first turns up after he has completed his experiment. He has taken no previous advice on how to design his experiment to test previously posed hypotheses. He has already browsed through his data in an attempt to "sort out the facts" and in this browsing may well have noticed something that "looks significant". The statistician is being consulted to verify that the significance is real. Can the experimenter reasonably expect the statistician to use data to test hypotheses which could well have been suggested by these data, and to carry out the same test as he would have employed in conducting tests of properly postulated hypotheses? The answer is clearly that he cannot. Is it possible with honesty to do anything for such experimenters? One answer is that we can at least provide a method for the guidance of the experimenter—possibly for the purposes of hypothesis-formulation for subsequent experimentation—but that he must expect to pay a

penalty, and sometimes a heavy penalty, for his break with scientific method. This is an extreme, rather obvious form of data-snooping. Unfortunately it comes in subtler forms. There is the well-intentioned experimenter who sets out to test quite a specific hypothesis and who, finding that his experiment is inconclusive in that he cannot reject this hypothesis, then looks at the data again in the hope that a significant effect may be spotted.

2.3. *Multiple-hypothesis testing*

There is little doubt that the only satisfactory treatment of a multiple-hypothesis testing problem is to regard it in terms of decision theory and to encourage the experimenter to attempt to specify an appropriate loss function. Statisticians have been particularly reluctant to develop this, probably because of the inherent difficulties of decision theory and the vagueness of the notion of loss function to the experimenter. The most successful attempt so far is probably that of Lehmann (1957a, b). The usual substitute for such a complete specification is to lay down some intuitive procedure for testing, and to examine its properties. All such procedures of multiple-hypothesis testing are to some extent *ad hoc*, and a proponent of one procedure attempts to make a case that his is less *ad hoc* than that of his rivals, or to justify it by the reasonableness of its properties. It is this unambitious aspect of examining the properties of the *ad hoc* procedure of confidence-region testing of many hypotheses that we undertake in this paper.

2.3.1. *Introspection of an accepted hypothesis*

Frequently an experimenter has under consideration a number of nested hypotheses $\omega_1, \omega_2, \ldots, \omega_k$, where $\omega_0 \supset \omega_1 \supset \omega_2 \supset \ldots \supset \omega_k$. A standard statistical procedure, common in regression and analysis of variance, is that sometimes termed "fitting-the-models"; see, for example, Fraser (1958, pp. 266, 304–308). First, we test ω_1 within ω_0 at the 5 per cent significance level, say. If ω_1 is rejected our analysis ends there; if it is not rejected we proceed to carry out some test of ω_2 within ω_1 (or possibly within ω_0), again at the 5 per cent level. If we reject ω_2 at this stage the analysis is terminated; otherwise it continues along the obvious path. We finally dismiss as unreasonable the first hypothesis rejected by this procedure. The interpretation of significance level is, for such procedures—as it is for all multiple-hypothesis testing—entangled in the web of conditions attaching to tests in the hierarchy of tests. What does it mean, in terms of liability to error, to reject ω_2 at the 5 per cent significance level when in fact we carry out this test only if a previous 5 per cent test has given a particular result? How sensible indeed is it to persist in the use of 5 per cent at every stage of the analysis? There is no blatant overthrow of scientific principle here, but in a sense there is a latent departure from it since our particular use of the data can hoodwink us into a kind of placid acceptance of a use and interpretation of significance levels which can hardly stand up to careful scrutiny.

Since statistical tests are designed more for the rejection of hypotheses than for their acceptance we note that it would be more appropriate to start our investigation with ω_k, the smallest region of the parameter space to come under scrutiny. We first ask if we can reject ω_k; if so, can we become more ambitious and regard ω_{k-1} as a candidate for rejection, and so on? This equivalent view of the problem really falls into the category of "enlarging rejected hypotheses" which we now consider informally.

2.3.2. *Enlarging rejected hypotheses*

It is common statistical practice, when a test rejects a hypothesis ω_1, to investigate wider hypotheses containing ω_1 in the hope that one or other of these may be also rejected. For example, in analysis-of-variance work, the partitioning of a significant sum of squares, associated with a hypothesis ω_1, into components is an attempt to explore with each component some hypothesis wider than ω_1; see, for example, Goulden (1952, p. 87). Similar action by partitioning is also taken in contingency-table analysis; see, for example, Rao (1952, pp. 192–196). Such follow-up tests are often carried out at the same significance level as the basic test and again we may question what interpretation, if any, can be placed on this significance level. Should we not be much more cautious in discarding these wider hypotheses than we are in our rejection of the smaller hypothesis first tested?

3. Data-snooping in a Linear Model

The extreme case of uninhibited data-snooping in a linear model sets the scene for subsequent analyses by providing most of their important ingredients while retaining mathematical simplicity.

We suppose that x is a vector observation on a $N(\theta, \sigma^2 I_n)$ random variable and that the experimenter and statistician are agreed that the θ-parameter space—which is the one of main interest, σ really being an unknown nuisance parameter—is not the whole of R^n, but some linear subspace ω_0 of R^n of dimension d_0. It is usual to refer to this basic parameter space as the model. This is the natural setting for any fixed-effects model; the choice of the linear space ω_0 is often simply dictated by the experimental design. We suppose that the experimenter has not made up his mind what hypothesis he wishes to test. Let $\theta^0(x)$ denote the maximum-likelihood or least-squares estimate of the true parameter value θ^*, say, so that $\theta^0(x)$ is the projection of x on ω_0. Then, denoting distance in R^n between x and y by $\|x-y\|$, we recall that $\|\theta^0(x)-\theta^*\|^2/(d_0 s^2)$ is an observation on an $F(d_0, n-d_0)$ distribution, where $s^2 = \|x-\theta^0(x)\|^2/(n-d_0)$. Hence it is natural, in the absence of any indication of interest in particular hypotheses, for us to take, as a neutralist confidence region of confidence coefficient $1-\alpha$, the region

$$C(x) = \{\theta : \|\theta-\theta^0(x)\|^2/(d_0 s^2) \leqslant F(d_0, n-d_0; 1-\alpha)\}, \tag{1}$$

where $F(\nu_1, \nu_2; p)$ denotes the p-probability point of an $F(\nu_1, \nu_2)$ distribution. If the experimenter suggests a linear hypothesis ω_1 of dimension $d_1 < d_0$, then we use $C(x)$ as the basis of a confidence-region test as defined in Section 1. The consequences of this are most readily assessed by the construction of the equivalent critical region $A(\omega_1)$ defined by

$$A(\omega_1) = \{x : \omega_1 \cap C(x) = \emptyset\}. \tag{2}$$

(Note that we do not here restrict ω_1 to be a homogeneous linear hypothesis; the set ω_1 is, of course, at least parallel to some linear subspace of ω_0 and is a subset of ω_0.) Now

$$\omega_1 \cap C(x) = \emptyset$$

if and only if, for all $\theta \in \omega_1$,

$$\|\theta-\theta^0(x)\|^2 > d_0 s^2 F(d_0, n-d_0; 1-\alpha),$$

or, equivalently,

$$\min_{\theta \in \omega_1} \|\theta-\theta^0(x)\|^2 > d_0 s^2 F(d_0, n-d_0; 1-\alpha).$$

Since $\theta^1(x)$, the least-squares estimate of θ^* on hypothesis ω_1, is the projection of $\theta^0(x)$ on ω_1 we have immediately that

$$\min_{\theta \in \omega_1} \| \theta - \theta^0(x) \|^2 = \| \theta^1(x) - \theta^0(x) \|^2,$$

and so

$$A(\omega_1) = \{x : \| \theta^1(x) - \theta^0(x) \|^2 > d_0 s^2 F(d_0, n-d_0; 1-\alpha)\} \tag{3}$$

$$= [x : \| \theta^1(x) - \theta^0(x) \|^2 / \{(d_0 - d_1) s^2\}$$

$$> \{d_0/(d_0-d_1)\} F(d_0, n-d_0; 1-\alpha)]. \tag{4}$$

A feature of critical region (4) is that it is defined by means of the usual statistic for testing ω_1 against $\omega_0 - \omega_1$ but that the critical value for comparison is

$$\{d_0/(d_0-d_1)\} F(d_0, n-d_0; 1-\alpha)$$

instead of the customary α significance value $F(d_0-d_1, n-d_0; 1-\alpha)$. The proposed method can thus be interpreted in terms of the penalty that it imposes on the data-snooper by reducing the significance level α (corresponding to confidence coefficient $1-\alpha$) to β, determined by

$$F(d_0-d_1, n-d_0; 1-\beta) = \{d_0/(d_0-d_1)\} F(d_0, n-d_0; 1-\alpha). \tag{5}$$

Table 1, which serves both the present Section and Section 5, gives the solution β of the equation

$$F(b, c; 1-\beta) = (a/b) F(a, c; 1-\alpha) \tag{6}$$

TABLE 1

Values of the equivalent significance level β corresponding to $\alpha = 0{\cdot}05$, for the linear case

c	a	b	β	c	a	b	β
2	2	1	0·025	40	10	9	0·034
	4	3	0·038			5	0·0040
		1	0·013			1	0·000049
5	5	4	0·034		20	19	0·039
		1	0·0040			10	0·0015
	10	9	0·041		40	39	0·044
		5	0·014			20	0·00052
		1	0·0010	100	10	9	0·034
10	5	4	0·031			5	0·0032
		1	0·0022		20	19	0·038
	10	9	0·038			10	0·00086
		5	0·0083		50	49	0·042
		1	0·00028			25	0·000085
	20	19	0·043		100	90	0·017
		10	0·0062			50	0·0000071
		1	0·000022				

for a selection of values a, b, c and for $\alpha = 0{\cdot}05$. Values of β are easily determined since

$$\beta = I_x(\tfrac{1}{2}c, \tfrac{1}{2}b) = 1 - I_{1-x}(\tfrac{1}{2}b, \tfrac{1}{2}c), \tag{7}$$

where $I_x(\frac{1}{2}c, \frac{1}{2}b)$ denotes an incomplete beta ratio (Pearson, 1934) and

$$x = c/\{c + aF(a, c;\ 1-\alpha)\}.$$

The equivalent significance level β of the confidence-region test of the linear hypothesis ω_1 is obtained when the table is entered at $a = d_0$, $b = d_0 - d_1$, $c = n - d_0$.

The application of confidence-region testing has been reduced, in (3) and (4), to a test which involves exactly the same basic computations as are required for standard tests of linear hypotheses. We are now in a position to indicate ways in which this procedure has intuitive appeal. By choosing $C(x)$ as confidence region we effectively regard as implausible any parameter value that falls outside $C(x)$, and so we would expect the confidence-region test based on $C(x)$ to reject such a parameter value, or simple hypothesis, at the significance level α corresponding to the confidence coefficient $1-\alpha$. This follows from the fact that a simple hypothesis has dimension $d_1 = 0$ so that $b = a$ and $\beta = \alpha$. As long as the experimenter confines himself to simple hypotheses he incurs no penalty. If, however, he becomes more ambitious and thinks he can reject a larger hypothesis ω_1 of higher dimension then we should proceed with more caution. The larger the region of the parameter space that is involved the more stringent we ought to make our assessment of the possibility of rejection. This is precisely the feature of confidence-region tests. If $d_1 > 0$ we have $b < a$ and so $\beta < \alpha$; the greater d_1 is, the greater is the difference between a and b and so the smaller β is, that is, the more reluctant we are to discard the hypothesis. It is clear from Table 1 that the equivalent significance level can be very small when the experimenter is really ambitious. A further property is that if $d_1 \ll d_0$ then β is approximately equal to α. This again seems very reasonable since, by attempting to discard a hypothesis of dimension d_1, the experimenter is not being very ambitious when the original model has high dimensionality.

We also note that, should the experimenter offer successively all the simple hypotheses in the parameter space, he would succeed in rejecting the whole of $\omega_0 - C(x)$. There is no anomaly here; by asking about the plausibility of each parameter value the experimenter turns his multiple-hypothesis data-snooping problem into the respectable one of estimation, and the statistician offers him a confidence region in return.

4. Data-snooping in a General Parametric Model

We suppose that the experiment has outcome space $\mathcal{X}$ and is described by a density function of the form $f(\cdot, \theta)$ on $\mathcal{X}$, where θ is d_0-dimensional with components that remain independent under the restriction $\theta \in \omega_0$. Corresponding to parameter value θ the information matrix B_θ of the experiment has (i,j) component

$$-E_\theta\{\partial^2 \log f(\cdot, \theta)/\partial\theta_i \partial\theta_j\},$$

where E_θ denotes expectation with respect to the density function $f(\cdot, \theta)$. We further suppose that we have available a set $x = (x_1, \ldots, x_n)$ of the outcomes of a large number n of independent replicates of the experiment, and that $\theta^0(x)$ is the maximum-likelihood estimate (under ω_0) of θ^*, the true parameter value. To deal with a data-snooping

experimenter we may reasonably take, as our asymptotic neutralist confidence region with coefficient $1-\alpha$, the region

$$C(x) = [\theta : n\{\theta - \theta^0(x)\}' B_{\theta^*(x)}\{\theta - \theta^0(x)\} \leqslant \chi^2(d_0; 1-\alpha)], \tag{8}$$

where $\chi^2(d; p)$ denotes the p-probability point of a chi-squared distribution with d degrees of freedom. If the experimenter then suggests a hypothesis ω_1, which places $d_0 - d_1$ functional restrictions on the components of the parameter, so effectively reducing the dimension of the parameter to d_1, we would use $C(x)$ as the basis of a confidence-region test. Our first problem here is again to devise a simple means of investigating whether or not $\omega_1 \cap C(x) = \emptyset$.

The fact that θ^0 is asymptotically distributed as $N\{\theta^*, (nB_{\theta^*})^{-1}\}$ can be made the basis of a development which exploits the asymptotic equivalence of the theory to that of a linear model with known covariance matrix; Lehmann (1959, pp. 303–311) indicates the appropriate technique. We omit the tedious details of the argument which leads to the result that, for large n,

$$\min_{\theta \in \omega_1} n(\theta - \theta^0)' B_{\theta^*}(\theta - \theta^0) = n(\theta^1 - \theta^0)' B_{\theta^*}(\theta^1 - \theta^0),$$

where θ^1 denotes the maximum-likelihood estimator under hypothesis ω_1. Now the statistic $n(\theta^1 - \theta^0)' B_{\theta^*}(\theta^1 - \theta^0)$ is asymptotically equivalent to any of the usual asymptotic test statistics for testing ω_1 within the model ω_0—the likelihood-ratio statistic T_{10}, the Lagrange-multiplier statistic V_{10} and the Wald statistic W_{10}; see Aitchison (1962) for the definition and notation of these test statistics. It follows that, for large n,

$$\min_{\theta \in \omega_1} n(\theta - \theta^0)' B_{\theta^*}(\theta - \theta^0) = T_{10}$$

and so, following an argument similar to that of Section 3, we have the critical region $A(\omega_1)$ of the confidence-region test of ω_1 given by

$$A(\omega_1) = \{x : T_{10}(x) \geqslant \chi^2(d_0; 1-\alpha)\}, \tag{9}$$

or one of the other two asymptotically equivalent forms.

We recall that a large-sample test of a properly postulated hypothesis ω_1 would use as critical region of size α

$$\{x : T_{10}(x) \geqslant \chi^2(d_0 - d_1; 1-\alpha)\}.$$

Thus we again see that the confidence-region test forces us to use the same test statistic as we would have used for a properly posed hypothesis, but at a different significance level. The equivalent significance level β is here given by

$$\chi^2(d_0 - d_1; 1-\beta) = \chi^2(d_0; 1-\alpha). \tag{10}$$

Table 2, which is also required for Section 6, gives the solution β of the equation

$$\chi^2(b; 1-\beta) = \chi^2(a; 1-\alpha) \tag{11}$$

for a selection of values of a, b and for $\alpha = 0{\cdot}05$. The solution is easily obtained as

$$\beta = 1 - I(u, p), \tag{12}$$

where $I(u,p)$ is the incomplete gamma ratio, as defined by Pearson (1922), with $p = \frac{1}{2}b - 1$, $u = \chi^2(a; 1-\alpha)/(2b)^{\frac{1}{2}}$.

TABLE 2

Values of the equivalent significance level β corresponding to $\alpha = 0{\cdot}05$, for the asymptotic case

a	b	β	a	b	β
2	1	0·014	20	19	0·036
				15	0·0078
5	4	0·026		10	0·00050
	3	0·011		5	0·0000078
	2	0·0040			
	1	0·00088	40	39	0·040
				30	0·0029
10	9	0·032		20	0·000032
	5	0·0026			
	1	0·000018			

The confidence-region test here displays all the features of the test of Section 3. For example, for small d_1 (for which a is near b) the experimenter is considering a small part of ω_0 and so β is not too far below α. If d_1 is large then Table 2 shows that the experimenter may have to pay a high price for being initially unsure of what the direction of his investigation should be.

5. MULTIPLE-HYPOTHESES TESTING IN A LINEAR MODEL

5.1. *The Basic Confidence-region Test*

The analyses, to which we now turn our attention, stem from one basic type of confidence-region test, which can conveniently be developed in the following setting. An experimenter, before he conducts his experiment, puts forward quite a specific hypothesis which he wishes to test. Once he learns the result of this test he may wish to consider other hypotheses. With this experimental setting in mind we look for a confidence-region test which provides the experimenter with a satisfactory test of his original hypothesis, and which serves as the basis of his subsequent tests. We suppose that the experiment is described by the linear model ω_0 of Section 3 and that the hypothesis of first interest is ω, a linear subspace of ω_0; a typical follow-up hypothesis, in the applications we consider, is a subspace ω_1 of ω_0, where ω_1 contains ω. The most convenient specification of ω_0, ω_1 and ω for our purposes is in terms of the null spaces of certain matrices. Since $\omega_0 \supset \omega_1 \supset \omega$ we can find a matrix $[H_0' H_1' H_2']$ with orthogonal columns such that

$$\omega_0 = \{\theta : H_0\theta = 0\},$$

$$\omega_1 = \omega_0 \cap \{\theta : H_1\theta = 0\},$$

$$\omega = \omega_1 \cap \{\theta : H_2\theta = 0\} = \omega_0 \cap \{\theta : H\theta = 0\},$$

where $H' = [H_1' H_2']$. If the dimensions of $\omega_0, \omega_1, \omega$ are d_0, d_1, d, then the matrices H_0, H_1, H_2, H are of full ranks $n-d_0$, d_0-d_1, d_1-d, d_0-d.

To test ω within the model ω_0 it seems reasonable to construct our confidence region by way of the restriction on $H\theta$. Noting that $H\theta^0$ is distributed as $N(H\theta, \sigma^2 I_{d_0-d})$, we see that the confidence region

$$C(x) = [\theta : \{H\theta - H\theta^0(x)\}' \{H\theta - H\theta^0(x)\} \leqslant s^2(d_0-d)\, F(d_0-d, n-d_0;\, 1-\alpha)], \tag{13}$$

where $s^2 = \|x-\theta^0(x)\|^2/(n-d_0)$ has confidence coefficient $1-\alpha$ and $C(x)$ seems a natural region on which to base our confidence-region test procedure. The region $C(x)$ is essentially a reinterpretation of the type of confidence ellipsoid used by Scheffé (1953; 1959, p. 69) as the basis of his S-method of multiple comparison. We have chosen to imbed our region in the full parameter space ω_0 rather than introduce the Scheffé concept of spaces of estimable functions—such as the set of linear combinations of the elements of $H\theta$—which are an unnecessary complication in our analysis. It is interesting to note that Scheffé (1959, p. 82) substantially suggests a type of confidence-region test based on spaces of estimable functions, but does not develop it along the lines studied here. We advocate in Sections 5.3 and 5.4 below more far-reaching applications of the confidence-region procedure than Scheffé envisages in his multiple-comparison technique.

For $C(x)$ to be a satisfactory region it is, of course, necessary that the test of ω within ω_0 based on it should be the standard F-test of size α. This property is verified by Scheffé; there is no need to give an independent proof here, but we shall note the result as a special case of our analysis of Section 5.2.

Note that if ω is accepted then $\theta \in C(x)$ and so any (homogeneous) linear hypothesis can never be rejected. This may seem harsh to the inveterate data-snooper but records little more than the fact that justice has been done. For the experimenter clearly stated that he pinned his hopes on rejecting ω, and expressed no interest in any other subspace of ω_0. It seems not unreasonable therefore that, on not being able to achieve this minimum ambition of rejecting ω, he should be placed in a position which forces him to accept for the time being the whole of ω_0. Data-snooping is not allowable subsequent to acceptance of a considered hypothesis.

If ω is rejected our attention is directed to the study of the confidence-region test of a larger hypothesis ω_1 and to the question of whether an equivalent significance level interpretation, on the lines of Section 3, is possible.

5.2. *Testing a Wider Hypothesis; Equivalent Significance Levels*

Our first task is to find the critical region

$$A(\omega_1) = \{x : \omega_1 \cap C(x) = \emptyset\}$$

associated with the confidence-region test of ω_1 within ω_0, based on $C(x)$. We show that

$$A(\omega_1) = [x : \|\theta^1(x) - \theta^0(x)\|^2 > s^2(d_0-d)\, F(d_0-d, n-d_0;\, 1-\alpha)]. \tag{14}$$

This result is readily established since

$$\omega_1 \cap C(x) = \emptyset$$

if and only if

$$\min_{\theta \in \omega_1} \{\theta - \theta^0(x)\}' H'H \{\theta - \theta^0(x)\} > s^2(d_0-d)\, F(d_0-d, n-d_0;\, 1-\alpha). \tag{15}$$

The fact that $H_1\theta = 0$ for $\theta \in \omega_1$ reduces the left side of (15) to

$$\{\theta^0(x)\}' H_1' H_1 \theta^0(x) + \min_{\theta \in \omega_1} \{\theta - \theta^0(x)\}' H_2' H_2 \{\theta - \theta(x)\}.$$

Now if $\theta^1(x)$ is the least-squares estimate of θ^* under ω_1, then $\theta^1(x) \in \omega_1$ and

$$\theta^1(x) - \theta^0(x) = H_1' H_1 x$$

so that

$$H_2\{\theta^1(x) - \theta^0(x)\} = 0.$$

Hence the minimum of zero is attained by the semi-positive definite form at $\theta^1(x)$, and so (14) follows since

$$H_1' H_1 \theta^0(x) = \theta^1(x) - \theta^0(x).$$

The region (14) can be expressed as

$$A(\omega_1) = [x : \| \theta^1(x) - \theta^0(x) \|^2 / \{s^2(d_0 - d_1)\} > \{(d_0 - d)/(d_0 - d_1)\} F(d_0 - d, n - d_0; 1 - \alpha)], \quad (16)$$

and so we have a situation similar to that of Section 3, for the confidence-region test uses the test statistic of the standard F test of ω_1 within ω_0 but compares it against the critical value

$$\{(d_0 - d)/(d_0 - d_1)\} F(d_0 - d, n - d_0; 1 - \alpha)$$

instead of the usual

$$F(d_0 - d_1, n - d_0; 1 - \alpha).$$

The test procedure is thus again very easy to apply. We can again define, as in Section 3, an equivalent significance level β related to α by (6), where $a = d_0 - d$, $b = d_0 - d_1$, $c = n - d_0$. Typical values of β can again be obtained from Table 1. For the special case where $\omega_1 = \omega$ we have $d_1 = d$ and $\beta = \alpha$, which is the Scheffé result that the confidence-region test is equivalent to the standard test as far as ω is concerned. We note here also that, in presenting his S-method, Scheffé (1959, Theorem of Section 3.5) is concerned essentially with hypotheses involving one restriction and so of dimension $d+1$. He prefers to retain a confidence-region form for his results rather than convert to equivalent critical regions, as we have done here.

The confidence-region test here displays features similar to those of Section 3. In particular if the experimenter wishes to test a hypothesis ω_1 of much greater dimension than ω then d_1 is appreciably greater than d, and β will be correspondingly much less than α. The test guards against over-ambition by applying more stringent criteria than the usual test.

5.3. *Introspection of Accepted Hypotheses; the Nested Method*

The logical procedure for the "introspection of accepted hypotheses" was indicated in Section 2 to be the examination in sequence of nested hypotheses

$$\omega_k \subset \omega_{k-1} \subset \ldots \subset \omega_1,$$

a wider hypothesis being examined only if its predecessor is rejected. The first test is therefore of ω_k within ω_0 and so we advocate the construction of a confidence region $C(x)$ as in Section 5.1 for this purpose; further we make $C(x)$ the basis of a confidence-region test procedure for testing all subsequent hypotheses. It seems sound

sense to build up our region of plausible values of θ relative to ω_k, for the rejection of ω_k is our least ambitious aim. With this procedure we obtain a standard test of ω_k but as we become more ambitious and try to reject wider hypotheses we become more stringent in our assessments, essentially adopting lower significance levels than the fixed-level tests often advocated.

5.4. *Enlarging a Rejected Hypothesis—Partitioning*

The analysis of Sections 5.1 and 5.2 allows us to compare the application of the commonly used fixed-level procedure and the confidence-region test procedure to the testing of wider hypotheses by partitioning of test statistics. As a typical study we consider how the rejection of ω is followed by the partitioning of the statistic which tested ω into two components for the separate testing of the larger hypotheses ω_1 and ω_2, where $\omega = \omega_1 \cap \omega_2$. Now for the partitioning to be possible in the analysis-of-variance setting, ω_1 and ω_2 must be orthogonal subspaces of ω_0. So, without loss of generality, we may take the hypotheses to be, in the notation of Section 5.1,

$$\omega_1 = \omega_0 \cap \{\theta : H_1 \theta = 0\}$$

as before, and

$$\omega_2 = \omega_0 \cap \{\theta : H_2 \theta = 0\},$$

with dimension d_2. If $\theta^0(x)$, $\theta^1(x)$, $\theta^2(x)$, $\theta(x)$ denote the least-squares estimates of θ^* under $\omega_0, \omega_1, \omega_2, \omega$, respectively, and $a = d_0 - d$, $a_1 = d_0 - d_1$, $a_2 = d_0 - d_2$, then the commonly used critical regions for testing ω, ω_1 and ω_2 within ω_0 are based respectively on the following inequalities:

$$T(x) \equiv \|\theta(x) - \theta^0(x)\|^2/s^2 > aF(a, n-d_0;\ 1-\alpha) = aF(a), \text{ say},$$

$$T_1(x) \equiv \|\theta^1(x) - \theta^0(x)\|^2/s^2 > a_1 F(a_1),$$

$$T_2(x) \equiv \|\theta^2(x) - \theta^0(x)\|^2/s^2 > a_2 F(a_2).$$

The orthogonality of ω_1 and ω_2 within ω_0 gives

$$T(x) = T_1(x) + T_2(x),$$

$$a = a_1 + a_2,$$

so that the three inequalities can be written as

$$T_1(x) + T_2(x) > (a_1 + a_2) F(a_1 + a_2), \tag{17}$$

$$T_1(x) > a_1 F(a_1), \tag{18}$$

$$T_2(x) > a_2 F(a_2). \tag{19}$$

Standard practice, if (17) holds so that ω is rejected, is to ask which of (18) and (19) holds. If (18) holds we enlarge the region of rejection from ω to ω_1, if (19) holds we enlarge the region of rejection from ω to ω_2, if both (18) and (19) hold the region is enlarged from ω to $\omega_1 \cup \omega_2$.

Now if

$$(a_1 + a_2) F(a_1 + a_2) > a_1 F(a_1) + a_2 F(a_2), \tag{20}$$

then inequality (17) implies at least one of the inequalities (18) and (19). While (20) does not hold for $n - d_0 > 2$, examination of tables of F percentage points shows that $(a_1 + a_2) F(a_1 + a_2)$ is never much smaller than $a_1 F(a_1) + a_2 F(a_2)$, so that the

structure of the above procedure almost guarantees that evidence which rejects ω will automatically reject one of the wider hypotheses ω_1 and ω_2. If there are *a priori* grounds for believing that rejection of ω can only arise through the rejection of ω_1 or ω_2 or both, then we cannot object to this feature, although we may ask if the purpose of the initial test of ω has been clearly thought out. Often, however, there are no such *a priori* grounds and indeed the testing of ω_1 and ω_2 may arise as an afterthought to the rejection of ω. How sensible is it then to use the above procedure in which enlargement of ω almost automatically follows its rejection? For a case where (20) holds, the question just posed can be strengthened by the removal of the word "almost". Such a case occurs when $a_1 = a_2 = 1$, $n - d_0 = 2$, for which, at the 5 per cent level,

$$(a_1 + a_2) F(a_1 + a_2) = 38,$$

$$a_1 F(a_1) + a_2 F(a_2) = 37.$$

The confidence-region test procedure uses a region based on testing ω and, corresponding to the three inequalities (17), (18) and (19), we have

$$T_1(x) + T_2(x) > aF(a), \tag{21}$$

$$T_1(x) > aF(a), \tag{22}$$

$$T_2(x) > aF(a). \tag{23}$$

Rejection of ω does not now imply automatic or nearly automatic rejection of one or other of ω_1 and ω_2, and again we have the feature, desirable in many practical situations, of much more stringent assessment of data when larger hypotheses are under consideration.

A complementary problem—when do standard tests of ω_1 and ω_2 "induce" a good test of $\omega = \omega_1 \cap \omega_2$—is the subject of an interesting paper by Darroch and Silvey (1963).

6. Multiple-hypotheses Testing in a General Parametric Model

The procedures in Section 5 have asymptotic analogues in the general parametric model of Section 3. These we now record briefly using a notation as similar as possible to that of Section 5.

The hypotheses ω_1, ω_2 and ω are conveniently defined in terms of the null spaces of vector functions h_1, h_2 and h, with domain ω_0, by

$$\omega_1 = \omega_0 \cap \{\theta : h_1(\theta) = 0\},$$

$$\omega_2 = \omega_0 \cap \{\theta : h_2(\theta) = 0\},$$

$$\omega = \omega_1 \cap \omega_2 = \omega_0 \cap \{\theta : h(\theta) = 0\},$$

where $h' = [h_1' h_2']$. We suppose that the dimensions of h_1, h_2 and h are $d_0 - d_1$, $d_0 - d_2$ and $d_0 - d$. The corresponding matrices H_1', H_2' and H', which consist of first-order partial derivatives of h_1, h_2 and h and which are of orders $(d_0 - d_1) \times d_0$, $(d_0 - d_2) \times d_0$ and $(d_0 - d) \times d_0$, are assumed to be of ranks $d_0 - d_1$, $d_0 - d_2$ and $d_0 - d$, respectively.

If the experimenter intends first to test ω within the general model ω_0 then the natural confidence region $C(x)$ on which to base a test procedure is defined in terms of the inequality

$$n[h(\theta) - h\{\theta^0(x)\}]' (H' B^{-1} H)^{-1}_{\theta^0(x)} [h(\theta) - h\{\theta^0(x)\}] \leqslant \chi^2(d_0 - d;\, 1 - \alpha), \tag{24}$$

where $\theta^0(x)$ denotes, as before, the maximum-likelihood estimate under ω_0. The test of ω based on $C(x)$ can be shown to be the standard Wald test of size α or one of its asymptotic equivalents. The confidence-region test of a wider hypothesis ω_1, say, can be shown to have critical region—in the terminology of Aitchison (1962)—

$$\{x : T_{10}(x) > \chi^2(d_0 - d; 1-\alpha)\} \tag{25}$$

as compared with the standard asymptotic test of size α with critical region

$$\{x : T_{10}(x) > \chi^2(d_0 - d_1; 1-\alpha)\}. \tag{26}$$

The equivalent significance level β is therefore given by (11), where $a = d_0 - d$, $b = d_0 - d_1$, and displays the kind of features discussed in Section 5.2; again Table 2 provides typical values.

The analysis of nested hypotheses follows the line of development of Section 5.3. The analysis of partitioning—which includes as a special case the partitioning of statistics associated with contingency tables—follows a pattern similar to that of Section 5.4. The orthogonality requirement of hypotheses ω_1 and ω_2 is replaced by a separability requirement (Aitchison, 1962) with the condition $H_1 H_2' = 0$ replaced by the condition

$$H_1 B^{-1} H_2' = 0 \quad (\theta \in \omega_1 \cap \omega_2).$$

The awkward feature of standard practice, indicated in Section 5.4, persists in its less acute form with $\chi^2(a_1; 1-\alpha)$, etc. replacing $a_1 F(a_1)$, etc. in the argument. In contrast the corresponding confidence-region test displays more caution as we move to wider hypotheses.

The fact that contingency tables and their associated asymptotic tests constitute a special case of the general parametric model and its tests (Silvey, 1959; Aitchison and Silvey, 1960) ensures that the Goodman (1964) method of multiple comparison in contingency tables can be derived from the above theory.

7. Discussion

The basic feature of a confidence-region test is that, once the experimenter has stated his objectives in terms of the testing of hypotheses and has obtained the result of his experiment, a set of plausible parameter values is chosen and made the basis of all judgements. This seems to the author a most reasonable, and, in his more optimistic moods, a highly desirable, procedure, which takes some account of the "dimensions" or "sizes" of the hypotheses under consideration.

The two types of confidence region used in this paper appear natural choices in the linear and asymptotic cases, but it must be admitted that a natural choice may be far from obvious in other cases. For example, what is the natural confidence region for testing the hypothesis $\theta_1 + \theta_2 - \theta_3 = 0$ and based on observations from each of three negative exponential distributions with mean parameters $\theta_1, \theta_2, \theta_3$? Again, the question of how, in practice, to choose the confidence coefficient $1-\alpha$ has been left unconsidered. Where the first hypothesis to be tested seems to be unnaturally restrictive, a choice of α greater than the customary 0·05 would be sensible, say 0·10 or 0·20. For example, in a polynomial regression situation in which it can be assumed that degree k is certainly adequate and ω_i is the hypothesis that a polynomial of degree $k-i$ will suffice, then clearly we may be so sure of rejecting ω_k as to risk a coefficient based on $\alpha = 0{\cdot}10$ or 0·20. A similar kind of point is made by Scheffé (1959, p. 71) in relation to his S-method of multiple comparisons.

We deliberately leave this practical question aside here. The aim of the paper has been, by providing an alternative to some standard procedures, to provoke statisticians into considering these existing procedures more critically. There are dangers in assuming that practices are of necessity soundly based just because they are long-established in practice. Even if this paper has done something to make clear the interrelations of hypotheses it will have achieved something, for even among experienced statisticians the author has found peculiar misconceptions. For example, it is not always realized that the components of a partitioned statistic are aimed at testing wider, and not narrower, hypotheses than the hypothesis tested by the original statistic!

We have also left aside the considerable problem of comparing confidence-region testing with other proposed procedures, such as those of Duncan (1955). Such comparisons would probably be more *ad hoc* than the procedures being compared.

ACKNOWLEDGEMENTS

The final version of this paper has benefited from helpful discussions which followed the presentation of earlier versions at seminars at the Universities of Manchester and Newcastle, and from the constructive criticism of my colleagues in Liverpool. I wish to express my thanks to Miss Diane Sculthorpe who prepared Tables 1 and 2.

REFERENCES

AITCHISON, J. (1962), "Large-sample restricted parametric tests", *J. R. statist. Soc.* B, **24**, 234–250.
—— and SILVEY, S. D. (1960), "Maximum-likelihood estimation and associated tests of significance", *J. R. statist. Soc.* B, **22**, 154–171.
BEALE, E. M. L. (1960), "Confidence regions in non-linear estimation", *J. R. statist. Soc.* B, **22**, 41–88.
DARROCH, J. N. and SILVEY, S. D. (1963), "On testing more than one hypothesis", *Ann. math. Statist.*, **34**, 555–567.
DUNCAN, D. B. (1955), "Multiple range and multiple *F* tests", *Biometrics*, **11**, 1–42.
FRASER, D. A. S. (1958), *Statistics, an Introduction.* New York: Wiley.
GOODMAN, L. A. (1964), "A note on simultaneous confidence limits for cross-product ratios", *J. R. statist. Soc.* B, **26**, 86–102.
GOULDEN, C. H. (1952), *Methods of Statistical Analysis.* New York: Wiley.
LEHMANN, E. L. (1957a), "A theory of some multiple decision problems, I", *Ann. math. Statist.*, **28**, 1–25.
—— (1957b), "A theory of some multiple decision problems, II", *Ann. math. Statist.*, **28**, 547–572.
—— (1959), *Testing Statistical Hypotheses.* New York: Wiley.
NEYMAN, J. (1937), "Outline of a theory of statistical estimation based on the classical theory of probability", *Phil. Trans. roy. Soc.*, **236**, 333–380.
PEARSON, K. (1922), *Tables of the Incomplete* Γ*-Function.* Cambridge University Press.
—— (1934). *Tables of the Incomplete Beta-Function.* Cambridge University Press.
RAO, C. R. (1952), *Advanced Statistical Methods in Biometric Research.* New York: Wiley.
SCHEFFÉ, H. (1953), "A method for judging all contrasts in the analysis of variance", *J. Amer. statist. Ass.*, **47**, 381–400.
—— (1959), *The Analysis of Variance.* New York: Wiley.
SILVEY, S. D. (1959), "The Lagrangian multiplier test", *Ann. math. Statist.*, **30**, 389–407.

AITCHISON, J. (1965)

Likelihood-ratio and confidence region tests

Reprinted from *J. R. Statist. Soc.* B27, 245-50

Likelihood-ratio and Confidence-region Tests

By J. Aitchison

University of Liverpool

[Received December 1964]

Summary

A previous paper (Aitchison, 1964) proposed a method of confidence-region testing for multiple-hypothesis situations and developed the theory for linear univariate situations and for the asymptotic parametric case. The present paper shows that the results for these situations are in fact special cases of a basic relation between confidence-region tests and likelihood-ratio tests. This relation allows the easy extension of confidence-region testing to multivariate analysis.

1. Introduction

The method of confidence-region testing, as defined by Aitchison (1964), involves the construction of a confidence region $C(\mathbf{x})$ of size $1-\alpha$ for a parameter $\boldsymbol{\theta}$, based on the experimental observation $\mathbf{x}$, with the subsequent rejection of any hypothesis ω_1 (subset ω_1 of the parameter space ω_0) if $\omega_1 \cap C(\mathbf{x}) = \emptyset$. The method when applied to problems of testing many hypotheses $\omega_1, \omega_2, \ldots$ was shown to lead, in the case of linear univariate analysis (regression analysis, analysis of variance, etc.) and in the asymptotic parametric case, to the use of the standard test statistic, such as an F or χ^2 test statistic, but against a critical value in general differing from that of a direct test (of size α) of the hypothesis. One interpretation of this is that the effective size of the confidence-region test based on $C(\mathbf{x})$ is less than, but simply related to, α. The purpose of this paper is twofold. First, we show that these results are in fact derivable from a general relation between the method of confidence-region testing and a class of likelihood-ratio test statistics. Secondly, we indicate briefly how the general theory applies to multivariate situations, and provide an illustrative example.

2. General Confidence-region Test Theory

Let $L(\mathbf{x}, \boldsymbol{\theta})$ be the value of the likelihood function for observation $\mathbf{x} \in \mathcal{X}$ and parameter $\boldsymbol{\theta} \in \omega_0$. In what follows it is to be understood that in addition to $\boldsymbol{\theta}$ there may be nuisance parameters involved and that all suprema are to be calculated over the product of the stated $\boldsymbol{\theta}$-set and the nuisance parameter set. The general result will concern the type of situation described in Aitchison (1964), where the experimenter's aim is first to test a basic hypothesis (his choice of basic confidence region reflecting this aim) and, in the event of rejection of the basic hypothesis, to follow up by investigation of other "wider" hypotheses. In what follows it will be useful to use both $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ to denote possible values of the parameters. We suppose, for convenience of theoretical discussion only, that the basic hypothesis has been expressed in the form

$$\omega(\mathbf{a}) = \{\boldsymbol{\tau}: \mathbf{h}(\boldsymbol{\tau}) = \mathbf{a}\},$$

with typical follow-up hypothesis

$$\omega_1(\mathbf{b}) = \{\boldsymbol{\tau}: \mathbf{h}_1(\boldsymbol{\tau}) = \mathbf{b}\},$$

where $\mathbf{h}' = [\mathbf{h}_1'\mathbf{h}_2']$, $\mathbf{h}$, $\mathbf{h}_1$ and $\mathbf{h}_2$ being possibly vector functions and $\mathbf{a}$ and $\mathbf{b}$ vector constants of appropriate dimensions. Note that if $\mathbf{a}$ is an extension of the vector $\mathbf{b}$ then $\omega(\mathbf{a}) \in \omega_1(\mathbf{b})$. We denote the likelihood-ratio test statistic for testing $\omega(\mathbf{a})$ within the model ω_0 by $\Lambda(\mathbf{x}; \mathbf{a})$, i.e.

$$\Lambda(\mathbf{x}; \mathbf{a}) = \sup_{\tau \in \omega_0} L(\mathbf{x}, \tau) \Big/ \sup_{\tau \in \omega(\mathbf{a})} L(\mathbf{x}, \tau).$$

Similarly we define $\Lambda_1(\mathbf{x}; \mathbf{b})$ as the likelihood-ratio test statistic for testing $\omega_1(\mathbf{b})$ within ω_0. The following lemma gives a basic relation between Λ_1 and Λ.

Lemma.

$$\inf_{\theta \in \omega_1(\mathbf{b})} \Lambda\{\mathbf{x}; \mathbf{h}(\theta)\} = \Lambda_1(\mathbf{x}; \mathbf{b}).$$

Proof.

$$\inf_{\theta \in \omega_1(\mathbf{b})} \Lambda\{\mathbf{x}; \mathbf{h}(\theta)\} = \sup_{\tau \in \omega_0} L(\mathbf{x}, \tau) \Big/ \sup_{\theta \in \omega_1(\mathbf{b})} \sup_{\tau \in \omega\{\mathbf{h}(\theta)\}} L(\mathbf{x}, \tau)$$

$$= \sup_{\tau \in \omega_0} L(\mathbf{x}, \tau) \Big/ \sup_{\tau \in \omega_1(\mathbf{b})} L(\mathbf{x}, \tau),$$

since

$$[\tau: \tau \in \omega\{\mathbf{h}(\theta)\}, \text{ where } \theta \in \omega_1(\mathbf{b})]$$

$$= [\tau: \mathbf{h}(\tau) = \mathbf{h}(\theta), \text{ where } \mathbf{h}_1(\theta) = \mathbf{b}]$$

$$= [\tau: \mathbf{h}_1(\tau) = \mathbf{b}]$$

$$= \omega_1(\mathbf{b}).$$

We now make an assumption about the class of likelihood-ratio tests involved, namely that the critical region of size α for testing $\omega(\mathbf{a})$ within ω_0 is

$$A\{\omega(\mathbf{a})\} = \{\mathbf{x}: \Lambda(\mathbf{x}; \mathbf{a}) > c_\alpha(\Lambda)\},$$

where the critical value $c_\alpha(\Lambda)$ is the *same for each* $\mathbf{a}$. A similar assumption is supposed to hold for likelihood-ratio tests of $\omega_1(\mathbf{b})$. If, moreover, the test for each $\omega(\mathbf{a})$ is of exact size α for each $\theta \in \omega(\mathbf{a})$ then we have, for each $\mathbf{a}$,

$$P[A\{\omega(\mathbf{a})\} \mid \theta] = \alpha \tag{1}$$

for every $\theta \in \omega(\mathbf{a})$, where $P(\cdot \mid \theta)$ is the probability measure associated with θ. As confidence region based on $\mathbf{x}$ and designed initially for the testing of hypotheses of type $\omega(\mathbf{a})$ we take

$$C(\mathbf{x}) = [\theta: \Lambda\{\mathbf{x}; \mathbf{h}(\theta)\} \leqslant c_\alpha(\Lambda)]. \tag{2}$$

We note that

$$P\{\mathbf{x}: C(\mathbf{x}) \ni \theta \mid \theta\} = P[\mathbf{x}: \Lambda\{\mathbf{x}; \mathbf{h}(\theta)\} \leqslant c_\alpha(\Lambda) \mid \theta]$$

$$= 1 - P(A[\omega\{\mathbf{h}(\theta)\}] \mid \theta)$$

$$= 1 - \alpha$$

for every $\theta \in \omega_0$ since $\theta \in \omega\{\mathbf{h}(\theta)\}$ for every $\theta \in \omega_0$. Thus the confidence region has exact size $1-\alpha$. If the tests are not of exact size so that (1) has to be replaced by

$$\sup_{\theta \in \omega(\mathbf{a})} P[A\{\omega(\mathbf{a})\} \mid \theta] = \alpha \tag{3}$$

then

$$P\{\mathbf{x}: C(\mathbf{x}) \ni \theta \mid \theta\} \geqslant 1-\alpha$$

for every $\theta \in \omega_0$, and so the confidence region defined by (2) has conservative size $1-\alpha$. In either case we have the following basic theorem.

Theorem. The confidence-region test of $\omega_1(\mathbf{b})$ based on $C(\mathbf{x})$ uses as critical region

$$A\{\omega_1(\mathbf{b})\} = \{\mathbf{x}: \Lambda_1(\mathbf{x}; \mathbf{b}) > c_\alpha(\Lambda)\}.$$

Proof. The confidence-region test of $\omega_1(\mathbf{b})$ based on $C(\mathbf{x})$ rejects $\omega_1(\mathbf{b})$ if and only if

$$\omega_1(\mathbf{b}) \cap C(\mathbf{x}) = \emptyset,$$

i.e. if and only if

$$\Lambda\{\mathbf{x}; \mathbf{h}(\theta)\} > c_\alpha(\Lambda)$$

for all $\theta \in \omega_1(\mathbf{b})$, i.e. if and only if

$$\inf_{\theta \in \omega_1(\mathbf{b})} \Lambda\{\mathbf{x}; \mathbf{h}(\theta)\} > c_\alpha(\Lambda).$$

The result of the theorem then follows by application of the Lemma.

The feature of this test is that it uses the customary likelihood-ratio test statistic Λ_1 for testing $\omega_1(\mathbf{b})$ within ω_0 but compares it against the critical value $c_\alpha(\Lambda)$ of the statistic Λ instead of the critical value $c_\alpha(\Lambda_1)$ appropriate to a test of size α. Such a simultaneous test procedure, which has also been arrived at from different considerations by Gabriel (1964), gives support to advocates of fixed-odds likelihood-ratio testing. Confidence-region tests of all hypotheses of the form $\omega(\mathbf{a})$ have size α but those of wider hypotheses such as $\omega_1(\mathbf{b})$ have their effective size β smaller than α and simply determined by

$$c_\beta(\Lambda_1) = c_\alpha(\Lambda). \tag{4}$$

The discussion of the properties of confidence-region procedures in Aitchison (1964, p. 468) thus generalizes.

3. Applications of the General Theory

Standard test statistics are usually simple monotonic functions of likelihood-ratio test statistics rather than these statistics themselves. In order to apply the general theory we have therefore to consider the consequences of using $U(\Lambda)$ and $U_1(\Lambda_1)$ instead of Λ and Λ_1; for the sake of definiteness we suppose that U and U_1 are monotonic increasing functions. Let $c_\alpha(\cdot)$ denote, as before, the upper α point of the stated statistic. It is then easy to see that, instead of comparing U_1 against $c_\alpha(U_1)$ as a direct test of size α would require, the confidence-region test of $\omega_1(\mathbf{b})$ compares U_1 against $U_1[U^{-1}\{c_\alpha(U)\}]$. Thus its equivalent size β is more readily determined by

$$c_\beta(U_1) = U_1[U^{-1}\{c_\alpha(U)\}] \tag{5}$$

than by (4). For example, in some applications we have U and U_1 of the form

$$\begin{aligned} U(\Lambda) &= k(\Lambda^l - 1), \\ U_1(\Lambda_1) &= k_1(\Lambda_1^l - 1); \end{aligned} \tag{6}$$

for such a case the inverse function U^{-1} is easily determined and (5) reduces to

$$c_\beta(U_1) = (k_1/k)\, c_\alpha(U). \tag{7}$$

The tests which arise in the confidence-region procedures considered by Aitchison (1964)—for linear hypotheses in univariate models and for parametric hypotheses in asymptotic situations—are all simple monotonic functions of likelihood-ratio test statistics, which satisfy the assumption of Section 2. Thus, these confidence-region procedures fall within the framework of this more general theory and the confidence regions $C(\mathbf{x})$ defined by (13) and (24) of the previous paper are in fact of the form (2) defined here and are of exact size $1-\alpha$. The advantage of the general result of confidence-region testing—that the customary test statistic is used but against a fixed critical value determined by the initial test—is the ease with which procedures are derived from it, for nothing more than a knowledge of the standard test statistic is required; there is no need to become involved in detailed investigation in special cases and it is not even necessary to find the basic $C(\mathbf{x})$ explicitly. This is of particular help in multivariate situations. For example, suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are independent p-dimensional observations on independent $N(\theta \mathbf{z}_1, \Sigma), \ldots, N(\theta \mathbf{z}_M, \Sigma)$ distributions, where θ is a $p \times q$ matrix parameter, Σ is a $p \times p$ covariance matrix (nuisance) parameter and the $\mathbf{z}_1, \ldots, \mathbf{z}_M$ are known constants. If $\omega(\mathbf{a}) = \{\theta : \theta \mathbf{H} = \mathbf{a}\}$, $\omega_1(\mathbf{b}) = (\theta : \theta \mathbf{H}_1 = \mathbf{b})$, where $\mathbf{H}$, of order $q \times r$, can be partitioned as $(\mathbf{H}_1, \mathbf{H}_2)$ with $\mathbf{H}_1$ of order $q \times r_1$, then the confidence-region test of $\omega_1(\mathbf{b})$ would use the usual likelihood-ratio test but instead of comparing against the upper α point of $U_{p,r_1,M-q}$ it compares against the upper α point of $U_{p,r,M-q}$; see Anderson (1958, Sections 8.3 and 8.4) for such tests and the definition of the U-distribution.

To provide a simple illustration of the use of such a procedure we consider the problem of assessing the effect of an air filter on the size distribution of particles in an *atmosphere*. In a common form of sampling, air is drawn through a sampler and particles in different size ranges are deposited at p (say) different *stages* in the sampler, the larger particles being at the earlier stages. An observation thus consists of a p-dimensional vector of quantities (or, to conform more closely to the usual model adopted, logarithms of quantities) of particles deposited at the various stages. The effect of introducing a filter is that the quantities of larger particles entering the atmosphere may be reduced. One way of measuring the effectiveness of the filter is thus to determine up to which stage (if any) there is a significant reduction in the quantities deposited. We suppose that altogether a set of m observations

$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$$

with, and a set of n observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ without, the filter are available. (In actual practice the experiment is more complicated than that described here because of the desirability of measuring the effect at different locations within the atmosphere. For illustrative purposes we have, however, ignored such complications.) The $m+n$ observations are independent, $\mathbf{x}_1, \ldots, \mathbf{x}_m$ being on a p-dimensional normal distribution $N(\alpha, \Sigma)$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n$ on $N(\beta, \Sigma)$.

Since interest is in the number of stages affected we require to examine differences between α and β through the system of nested hypotheses $\omega_1 \supset \omega_2 \supset \ldots \supset \omega_p = \omega$, where

$$\omega_i = \{(\alpha, \beta) : \alpha_i - \beta_i = 0\},$$

α_i and β_i being the subvectors of α and β consisting of their final i components; note that $\alpha_p = \alpha, \beta_p = \beta$. In terms of the general multivariate model introduced

earlier in this Section, we have $M = m+n$, $\theta = [\alpha, \beta]$, $q = 2$, $\mathbf{z}_1 = \ldots = \mathbf{z}_m = [1\ 0]'$, $\mathbf{z}_{m+1} = \ldots = \mathbf{z}_{m+n} = [0\ 1]'$, with

$$\mathbf{H} = \begin{bmatrix} 1 \ldots & 1 \\ -1 & -1 \end{bmatrix}$$

of order $2 \times p$ and $\mathbf{H}_i$, the "$\mathbf{H}_1$" corresponding to ω_i, consisting of i columns of $\mathbf{H}$; also $\mathbf{a} = \mathbf{0}, \mathbf{b} = \mathbf{0}$.

The confidence-region procedure is now easily described. Let

$$m\bar{\mathbf{x}} = \Sigma \mathbf{x}_j, \quad n\bar{\mathbf{y}} = \Sigma \mathbf{y}_j,$$

$$(m+n-2)\mathbf{S} = \Sigma(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + \Sigma(\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})',$$

and denote by $\bar{\mathbf{x}}_i, \bar{\mathbf{y}}_i$ the subvectors of $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ consisting of their final i components, and by $\mathbf{S}_i$ the $i \times i$ submatrix of $\mathbf{S}$ consisting of the components common to its final i rows and i columns. From standard multivariate theory (see, for example, Anderson, 1958, Section 5.3.3) the test of null hypothesis ω_i at significance level α uses critical region based on the inequality

$$U_i(\mathbf{x}, \mathbf{y}) = \frac{m+n-i-1}{i} \frac{mn(m+n-2)}{m+n} (\bar{\mathbf{x}}_i - \bar{\mathbf{y}}_i)' \mathbf{S}_i^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{y}}_i) > c_\alpha(F_{i,m+n-i-1}).$$

The relation of the test statistic U_i on the left side of this inequality to the corresponding likelihood-ratio test statistic Λ_i is

$$U_i(\Lambda_i) = \{(m+n-i-1)/i\}(\Lambda_i^{2/(m+n)} - 1),$$

which is of form (6). The confidence-region procedure takes as its first target the smallest hypothesis ω_p, which is thus the ω of the general theory. For ω_p we thus get the usual critical region of size α, rejecting ω_p (and thus deciding that at least the first stage is affected) if

$$U_p(\mathbf{x}, \mathbf{y}) > c_\alpha(F_{p,m+n-p-1}).$$

If we reject ω_p we proceed to examine successively the inequalities

$$U_i(\mathbf{x}, \mathbf{y}) > c_\alpha(F_{p,m+n-p-1}) \quad (i = p-1, p-2, \ldots, 1),$$

rejecting ω_i if the corresponding inequality is satisfied. Let ω_j be the last hypothesis to be rejected; note that the confidence-region procedure then guarantees the acceptance of all of $\omega_{j-1}, \ldots, \omega_1$. The conclusion is then that the first $p-j+1$ stages, and only those stages, are affected.

A computational point worth noting is that the $\mathbf{S}_i^{-1}$ can be generated in the process of inverting $\mathbf{S}$ provided one of the bordering techniques of matrix inversion is used.

It is of interest to consider the equivalent significance level β_i of the confidence-region test of ω_i. From (7) this is determined by

$$c_{\beta_i}(F_{i,m+n-i-1}) = \frac{(m+n-i-1)p}{(m+n-p-1)i} c_\alpha(F_{p,m+n-p-1}) \qquad (8)$$

so that

$$\beta_i = I_x\{\tfrac{1}{2}(m+n-i-1), \tfrac{1}{2}i\},$$

where I_x denotes an incomplete beta ratio (Pearson, 1934) and

$$x = (m+n-p-1)/\{m+n-p-1+pc_\alpha(F_{p,m+n-p-1})\};$$

see Aitchison (1964, equation (6)) for a slightly more special form of (8). Table 1 gives an example of such equivalent significance levels for the case of 6 stages and for two confidence region sizes, 0·90 and 0·75.

TABLE 1

Equivalent significance levels, β_i, for confidence region tests of ω_i for the case $p = 6$ and for $\alpha = 0{\cdot}10$ and 0·25

i	$\alpha = 0{\cdot}10$	$\alpha = 0{\cdot}25$
6	0·10	0·25
5	0·052	0·15
4	0·024	0·082
3	0·0091	0·037
2	0·0027	0·013
1	0·00049	0·0028

The table illustrates once more the added caution employed by confidence-region testing when the hypothesis under test widens, as compared with fixed-level testing. In the present practical situation we may be disinclined to overstate too readily the efficiency of the filter, and to meet this disinclination the confidence-region procedure seems well suited. Although equivalent significance levels are a poor substitute for a well-developed loss function they may at least prove a satisfactory guide for some statisticians and comparison of them for various α may provide a useful basis for the actual choice of α in a practical application.

REFERENCES

AITCHISON, J. (1964), "Confidence-region tests", *J. R. statist. Soc.* B, 26, 462–476.
ANDERSON, T. W. (1958), *An Introduction to Multivariate Analysis.* New York: Wiley.
GABRIEL, K. R. (1964), "Simultaneous test procedures" (Abstract), *Ann. math. Statist.*, 35, 1400.
PEARSON, K. (1934), *Tables of the Incomplete Beta-Function.* Cambridge University Press.

5 CONSTRUCTION OF F-OPTIMAL DESIGNS

A well-known result in the design of randomised controlled trials to compare a control against s different treatments is that in order to minimise the variance of a control versus treatment contrast we must assign $\sqrt{s}$ as many experimental units to the control as to each treatment. Motivated by a consultative problem in which interest is in a much more complex set of contrasts Aitchison (4:1961) sets out to investigate the nature of generalisations of the familiar result to a general set of contrasts. The main features of the approach are briefly as follows.

1. A definition of F-optimal designs which maximise the infimum of the power of the appropriate F-test over a region defined by the specified contrasts.
2. A proof that an F-optimal design has a minmax property in relation to the set of variances of the contrast estimators.
3. A demonstration that a direct approach to this maxinf or minmax problem is usually difficult.
4. The conversion of the problem, through game theory, to the computationally simpler problem of maximising a concave function of the form

$$\sum_{j=1}^{s+1} \left(\sum_{i=1}^{s} a_{ij} x_i \right)^{\frac{1}{2}} \qquad (x \in S^s)$$

over the s-dimensional simplex S^s, where a_{ij} $(i = 1,\ldots,s;\ j = 1,\ldots,s+1)$ are given non-negative constants.

5. The provision of a computational scheme for this maximisation problem and some examples of its application.

Apart from work on optimal design in polynomial regression most approaches to experimental design in 1961 were content with establishing or denying optimum properties of well-established designs. An interesting aspect of the above approach, therefore, is its provision of a computational construction for optimal designs.

AITCHISON, J. (1961)

The construction of optimal designs for the one-way classification analysis of variance

Reprinted from *J. R. Statist. Soc.* B23, 352-67

The Construction of Optimal Designs for the One-way Classification Analysis of Variance

By JOHN AITCHISON
University of Glasgow

[Received August 1960. Revised February 1961]

SUMMARY

A consideration of the purpose of some one-way classification experiments leads us to introduce a criterion of *F*-optimality of design, a compromise measure which concentrates on the detection of certain specified effects while allowing at least the inspection of a wider class of effects. To give added validity to this notion we first establish its equivalence to a min–max weighted variance design criterion. The actual construction of *F*-optimal designs is our main purpose. The direct min–max approach is difficult, and so our next main result directs attention to the equivalent problem of finding optimal strategies for a certain two-person zero-sum non-matrix game. Finally, we provide a computational scheme for the solution of this game and illustrate the technique by examples.

1. INTRODUCTION

IN a one-way classification experiment involving t treatments and using a fixed total number n of observations, a design problem arises when it is possible to assign in different ways the numbers n_j $(j = 1, 2, ..., t)$ of observations to be taken on the t treatments. We suppose that any t-way partition $(n_1 n_2 ... n_t)$ of the integer n is a possible *design*, allocating n_j observations to the jth treatment; and that the corresponding fixed-effects analysis-of-variance model concerns n independent random variables Z_{jk}, where Z_{jk} is $N(\theta_j, \sigma^2)$ $(j = 1, 2, ..., t; k = 1, 2, ..., n_j)$, with σ^2 unknown. The modifications to the theory for the case of σ^2 known are trivial and we prefer to treat the more realistic case. Any choice of design, by way of some criterion of optimality, must take account of the purpose of the experiment. Examination of the following examples provides the motivation for the introduction of the criterion of *F*-optimality in section 2. It is then the purpose of the paper to show how *F*-optimal designs may be constructed.

Example 1. This is the familiar case of comparing a control (which we take as treatment 1) with $s = t-1$ treatments. We are interested only in contrasts in the θ_j $(j = 1, 2, ..., t)$, and primarily in the special contrasts $\theta_1 - \theta_j$ $(j = 2, ..., t)$, i.e. the s simple control versus treatment contrasts. Our attitude to the experiment may be that we wish to test the null hypothesis that all contrasts are zero, just in case some unspecified contrast is of importance, while ensuring that the design we choose enables us as efficiently as possible to detect any significant control versus treatment contrasts. If we regard each of the set of specified contrasts as of equal importance, we may decide to choose the design which gives the *F*-test of zero contrasts the best chance of detecting when one or other of the inequalities $|\theta_1 - \theta_j| \geqslant \Delta$ $(j = 2, ..., t)$ is satisfied. We shall see that such a procedure is equivalent to the usual method of choosing the design which minimizes the variance of a control versus treatment

contrast estimator, and assigns $\sqrt{s}$ as many observations to the control as to each treatment.

Example 2. This is one of the many possible generalizations of Example 1. Suppose that we have t_1 standard, but untested, treatments or controls (numbered $1, 2, \ldots, t_1$) already in use and that $t_2 = t - t_1$ new treatments are proposed. We again take the precaution of being interested in all contrasts, but particularly in control versus control contrasts $\theta_j - \theta_{j'}$ $(j, j' = 1, 2, \ldots, t_1; j \neq j')$ and in control versus treatment contrasts $\theta_j - \theta_k$ $(j = 1, 2, \ldots, t_1; k = t_1+1, t_1+2, \ldots, t)$. We may then consider using the F-test of zero contrasts and choose our design to make as sure as possible that we detect whether one or other of the $\frac{1}{2}t_1(t+t_2-1)$ inequalities

$$|\theta_j - \theta_{j'}| \geq \Delta_1 \quad (j, j' = 1, 2, \ldots, t_1; j \neq j'),$$
$$|\theta_j - \theta_k| \geq \Delta_2 \quad (j = 1, 2, \ldots, t_1; k = t_1+1, \ldots, t)$$

holds.

Example 3. In six locations equal quantities of a perishable commodity are stored. A preservative is suggested and an experiment is proposed to measure the deterioration at each location of a number of treated and of untreated specimens. We consider the 12 location-treatment combinations as our 12 treatments, treatment j $(j = 1, 2, \ldots, 6)$ being the use of preservative in location j, and treatment $j+6$ $(j = 1, 2, \ldots, 6)$ being the use of no preservative in location j. We may then be interested only in whether the preservative is effective, i.e. only in the set of treatment (given location) contrasts $\theta_{j+6} - \theta_j$ $(j = 1, 2, \ldots, 6)$ and intend to use an over-all test of these; while, for some such reasons as differential costs in large-scale treatment in different locations, special importance attached to certain locations by management, etc., we wish to choose our design so that we readily detect when one or other of the inequalities

$$\tfrac{1}{3}\{(\theta_7 - \theta_1) + (\theta_8 - \theta_2) + (\theta_9 - \theta_3)\} - \tfrac{1}{2}\{(\theta_{10} - \theta_4) + (\theta_{11} - \theta_5)\} \geq \Delta_1,$$
$$\tfrac{1}{2}\{(\theta_{10} - \theta_4) + (\theta_{11} - \theta_5)\} - (\theta_{12} - \theta_6) \geq \Delta_2,$$
$$\theta_{12} - \theta_6 \geq \Delta_3$$

holds.

Such considerations of design are compromises between allowing for the unexpected, and so leaving open the possibility of following up the F-test with a data-snooping technique such as the Scheffé (1953) S-method, and making as sure as possible that certain specific inequalities in the parameters are detected. That such an approach is a reasonable one is supported by the analysis of section 2, where our main result shows its equivalence to the choice of a design by a min–max weighted variance principle.

Although much attention has been given to the problem of optimal design in linear experiments during the past decade, see the excellent survey by Elfving (1959), work has been concentrated on establishing or denying optimum properties of well-established designs, or on supplying sufficient conditions for designs to possess these optimum properties. Few computational methods for the actual construction of optimal designs were considered until very recently. An important breakthrough has been made for polynomial regression, where Guest (1958), Hoel (1958), Kiefer and Wolfowitz (1959) and Williams (1958) have constructed optimal designs under various criteria of optimality. Apart from this the outstanding contribution to computational techniques is the work of Kiefer and Wolfowitz (1959), who exploit a method based on Chebyshev approximation for the case of minimum variance estimation of a single parameter, and, by drawing analogies with the theory of a certain

type of game, point one direction of forward progress for more general problems. Our own problem leads, via a different route, to a solution which is most readily constructed by appealing to another type of game.

2. The Design Problem

Although in this paper we use the results of this section only for the special case of a one-way classification analysis of variance, it is an advantage, and adds no extra difficulty, to establish them in the wider setting of the general linear fixed-effects model for which they are valid. In particular, this allows us to indicate the special place of the one-way classification on the ladder of computational difficulty.

A linear fixed-effects model with parameter vector $\boldsymbol{\theta} = [\theta_1\ \theta_2 \dots \theta_t]'$ is given. If we use design d with associated real design matrix $\mathbf{M}_d$ of order $n \times t$, where n is a fixed integer, we make a vector-observation on the n-dimensional vector-valued random variable $\mathbf{Z}$, which is $N(\mathbf{M}_d\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$, i.e. we make n real observations on each of the independent components of $\mathbf{Z}$. We assume that we are interested only in $r \leqslant t$ linearly independent combinations $\boldsymbol{\psi} = [\psi_1\ \psi_2 \dots \psi_r]' = \mathbf{G}\boldsymbol{\theta}$ of the parameters θ_j and so we confine attention to the class D_ψ of designs (with associated design matrices of order $n \times t$) for which $\boldsymbol{\psi}$ is estimable. Let $\beta(\boldsymbol{\psi}, d)$ denote the power at $\boldsymbol{\psi}$ of the F-test, using design $d \in D_\psi$, of $H_0(\boldsymbol{\psi} = \mathbf{0})$ versus $\bar{H}_0(\boldsymbol{\psi} \neq \mathbf{0})$; the size of the test and σ^2 are unimportant in what follows and so we omit them from our notation in order to simplify it. Let Ψ be any region in R^r, r-dimensional real space.

Definition of $F(\boldsymbol{\psi}, \Psi)$ optimal designs. A design $d^* \in D_\psi$ is said to be $F(\boldsymbol{\psi}, \Psi)$ optimal if

$$\inf_{\boldsymbol{\psi} \in \Psi} \beta(\boldsymbol{\psi}, d^*) = \max_{d \in D_\psi} \inf_{\boldsymbol{\psi} \in \Psi} \beta(\boldsymbol{\psi}, d),$$

i.e. if it maximizes the infimum in the region Ψ of the power of the F-test of the null hypothesis that $\boldsymbol{\psi} = \mathbf{0}$.

We are concerned here only with a very special form of Ψ. The examples of section 1 show that our main interest is often in special combinations of the ψ_j, say $\mathbf{h}_i'\boldsymbol{\psi}$ $(i = 1, 2, \dots, s)$ and whether or not one or other of the inequalities $\mathbf{h}_i'\boldsymbol{\psi} \geqslant \Delta_i > 0$ $(i = 1, 2, \dots, s)$ is satisfied. Thus throughout this paper we shall always take $\Psi = \cup \Psi_i$, where $\Psi_i = \{\boldsymbol{\psi} \in R^r : \mathbf{h}_i'\boldsymbol{\psi} \geqslant \Delta_i\}$. The relations

$$\begin{aligned} \{\boldsymbol{\psi} : \mathbf{h}'\boldsymbol{\psi} \leqslant -\Delta\} &= \{\boldsymbol{\psi} : -\mathbf{h}'\boldsymbol{\psi} \geqslant \Delta\}, \\ \{\boldsymbol{\psi} : |\mathbf{h}'\boldsymbol{\psi}| \geqslant \Delta\} &= \{\boldsymbol{\psi} : \mathbf{h}'\boldsymbol{\psi} \geqslant \Delta\} \cup \{\boldsymbol{\psi} : -\mathbf{h}'\boldsymbol{\psi} \geqslant \Delta\} \end{aligned} \tag{2.1}$$

show that we lose no generality by considering only the one-sided type of linear inequality above. The main result, Theorem 1, of this section is to establish the equivalence of the $F(\boldsymbol{\psi}, \Psi)$ principle and a type of min–max weighted variance principle of design choice. For this we require the following lemma.

Lemma 1. If $\mathbf{P}$ is a positive definite $r \times r$ matrix then

$$\min_{\boldsymbol{\psi} \in \Psi_1} \boldsymbol{\psi}'\mathbf{P}\boldsymbol{\psi} = \Delta_1^2/(\mathbf{h}_1'\mathbf{P}^{-1}\mathbf{h}_1).$$

Proof. The function $\boldsymbol{\psi}'\mathbf{P}\boldsymbol{\psi}$ is strictly convex in $\boldsymbol{\psi}$ in R^r. Hence

$$\begin{aligned} \min_{\boldsymbol{\psi} \in \Psi_1} \boldsymbol{\psi}'\mathbf{P}\boldsymbol{\psi} &= \min_{\mathbf{h}_1'\boldsymbol{\psi} = \Delta_1} \boldsymbol{\psi}'\mathbf{P}\boldsymbol{\psi} \\ &= \Delta_1^2/(\mathbf{h}_1'\mathbf{P}^{-1}\mathbf{h}_1), \end{aligned}$$

by a simple application of the Lagrange-multiplier technique.

Corollary. We have that

$$\min_{\psi \in \Psi} \psi' \mathbf{P} \psi = \min_{i} \Delta_i^2 / (\mathbf{h}_i' \mathbf{P}^{-1} \mathbf{h}_i).$$

Theorem 1. Let $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \dots \mathbf{h}_s]'$ and $\hat{\psi}_d$ denote the least-squares estimator of ψ for design $d \in D_\psi$. The following three statements are then equivalent.

(i) d^* is $F(\psi, \Psi)$ optimal.

(ii) d^* is $F(\mathbf{H}\psi, \Psi)$ optimal.

(iii) $$\max_i \operatorname{var}(\mathbf{h}_i' \hat{\psi}_{d^*}) / \Delta_i^2 = \min_{d \in D_\psi} \max_i \operatorname{var}(\mathbf{h}_i' \hat{\psi}_d) / \Delta_i^2.$$

Proof. We establish first the equivalence of (i) and (iii). The quantity $\beta(\psi, d)$ depends on ψ and d only in the form $\psi' \mathbf{V}^{-1}(\hat{\psi}_d) \psi$, where $\mathbf{V}(.)$ denotes "variance–covariance matrix of"; and is indeed a strictly monotonic increasing function of $\psi' \mathbf{V}^{-1}(\hat{\psi}_d) \psi$. Hence d^* is $F(\psi, \Psi)$ optimal if and only if

$$\min_{\psi \in \Psi} \psi' \mathbf{V}^{-1}(\hat{\psi}_{d^*}) \psi = \max_{d \in D_\psi} \min_{\psi \in \Psi} \psi' \mathbf{V}^{-1}(\hat{\psi}_d) \psi,$$

i.e. by the corollary to Lemma 1, if and only if

$$\min_i \Delta_i^2 / \{\mathbf{h}_i' \mathbf{V}(\hat{\psi}_{d^*}) \mathbf{h}_i\} = \max_{d \in D_\psi} \min_i \Delta_i^2 / \{\mathbf{h}_i' \mathbf{V}(\hat{\psi}_d) \mathbf{h}_i\},$$

i.e. if and only if

$$\max_i \mathbf{h}_i' \mathbf{V}(\hat{\psi}_{d^*}) \mathbf{h}_i / \Delta_i^2 = \min_{d \in D_\psi} \max_i \mathbf{h}_i' \mathbf{V}(\hat{\psi}_d) \mathbf{h}_i / \Delta_i^2.$$

The equivalence of (i) and (iii) follows on noting that

$$\operatorname{var}(\mathbf{h}_i' \hat{\psi}_d) = \mathbf{h}_i' \mathbf{V}(\hat{\psi}_d) \mathbf{h}_i.$$

If in the equivalence just established we replace (ψ, Ψ) by (ϕ, Φ), where $\phi = \mathbf{H}\psi$ and

$$\Phi = \cup_i \{\phi \in R^s : \phi_i \geqslant \Delta_i\},$$

we obtain the equivalence of (ii) and (iii) immediately on noting that d^* is $F(\mathbf{H}\psi, \Psi)$ optimal if and only if d^* is $F(\Phi, \Phi)$ optimal, and that $\hat{\phi}_d = \mathbf{H}\hat{\psi}_d$. This completes the proof.

Kiefer (1958) uses implicitly the notion of F-optimality and establishes its equivalence, with a Ψ different from ours, to another type of optimality involving the characteristic roots of $V(\hat{\psi}_d)$, but rejects the notion on the grounds that it assumes the use of the F-test, whereas a pure max–min power principle would not insist on this. Our own view is that there are many situations, as indicated in section 1, where there is a two-fold purpose of the experiment: to keep an eye on all ψ, which requires the use of an F-test, while focusing special attention on $\mathbf{H}\psi$. For those who remain suspicious of such a compromise procedure as F-optimality the results of Theorem 1 may clothe the idea in respectability. It is comforting to know, from the equivalence of (i) and (ii), that it makes no difference, as far as the choice of design is concerned, whether we initially carry out an F-test of the null hypothesis that $\psi = 0$ or restrict ourselves to the null hypothesis that $\mathbf{H}\psi = 0$, involving only the special parametric functions. Statement (iii) means that d^* is chosen to minimize the maximum weighted

(the weighting factor being $1/\Delta_i^2$) variance of the least-squares estimators of the s special parametric functions. If we had adopted an estimation or confidence-interval approach for these s special parametric functions, then an appropriate design criterion might just have been this min–max weighted variance one above, the weighting indicating that we regard the special parametric functions as not of equal importance. Such a procedure ensures that we minimize the maximum expected length of the usual confidence intervals, taking into account their possibly differing importance.

The simplification in the computation of d^* that we obtain in the one-way classification analysis of variance arises from the fact that $\text{var}(\mathbf{h}'\hat{\boldsymbol{\psi}}_d)$ is then a homogeneous linear function of $(1/n_1, 1/n_2, \ldots, 1/n_t)$ with non-negative coefficients. We are thus led to the problem of determining a design $\mathbf{n}^* = (n_1^* n_2^* \ldots n_t^*)$ such that

$$\max_i \sum_{j=1}^{t} a_{ij}/n_j^* = \min_{\mathbf{n}\in N} \max_i \sum_{j=1}^{t} a_{ij}/n_j,$$

where N is the set of all t-way partitions of the integer n, and $a_{ij} \geqslant 0$ $(i = 1, 2, \ldots, s;$ $j = 1, 2, \ldots, t)$. Clearly we have that, for each i, $a_{ij} > 0$ for some j. Moreover, we can assume that, for each j, $a_{ij} > 0$ for some i; for otherwise, if $a_{ij} = 0$ for all i, we would obviously take no observations on the jth treatment and so the problem would reduce to one involving at most $t-1$ treatments. It is also worth noting here that when an inequality of the form $|\mathbf{h}'\boldsymbol{\psi}| \geqslant \Delta$ is involved the breakdown by (2.1) leads to the consideration of only one expression of the form $\Sigma a_{ij}/n_j$, since

$$\text{var}(\mathbf{h}'\hat{\boldsymbol{\psi}}_d) = \text{var}(-\mathbf{h}'\hat{\boldsymbol{\psi}}_d).$$

Instead of considering a design $\mathbf{n}$, we may consider equivalently $\mathbf{y} = (y_1 y_2 \ldots y_t)$ where $y_j = n_j/n$ denotes the proportion of observations assigned to the jth treatment, and $\Sigma y_j = 1$. For the sake of mathematical simplicity we then adopt the following customary, though not fully justified, device of approximating to the problem involving discrete $\mathbf{y}$, by converting it to one of continuous variation. We allow $\mathbf{y}$ to vary continuously, stipulating only that

$$\mathbf{y} \in T = \left\{ \mathbf{y} \in R^t : y_j \geqslant 0 \quad (j = 1, 2, \ldots, t), \quad \sum_j y_j = 1 \right\},$$

the simplex in R^t. An optimal design $\mathbf{y}^*$ then satisfies

$$\max_i \sum_j (a_{ij}/y_j^*) = \min_{\mathbf{y}\in T} \max_i \sum_j (a_{ij}/y_j). \tag{2.2}$$

If $n\mathbf{y}^*$ turns out to be a vector with integral elements, as can be the case, then the original discrete problem is solved. Otherwise we hope to arrive, by rounding the y_j^* to nearer multiples of $1/n$, or setting $y_j^* = 1/n$ if $y_j^* < 1/n$, at a solution sufficiently close to the discrete optimum. If n is large we could justify this procedure by using the fact that there is then a $\mathbf{y}$ of the discrete-variation problem very close to any given member of T. For moderate n the procedure is more one of intuition and practical expediency; we can console ourselves, however, as do Kiefer and Wolfowitz (1959) when faced with a similar difficulty, that there is practical sense in providing a solution $\mathbf{y}^*$ (independent of n) of (2.2), valid for all large n and probably for moderate n in many situations, rather than in facing a major computational problem for each moderate n or abandoning the problem altogether.

From now on we consider the problem of finding a $\mathbf{y}^*$ satisfying (2.2) as *the* design problem. It is convenient to have some jargon for the quantity defined by

(2.2), which we term the *min–max* of the design problem. From the min–max of the design problem we may easily determine the max–min power in the definition of *F*-optimality. Bounds may be immediately placed on the min–max of the design problem by noting that

$$\max_i \left(\sum_j a_{ij}^2\right)^2 = \max_i \min_{y \in T} \sum_j a_{ij}/y_j \leqslant \text{min–max} \leqslant t \max_i \sum_j a_{ij}. \tag{2.3}$$

In the next section we make a direct approach to the min–max problem so that we may later high-light the efficiency of the game-theoretic approach.

3. A Direct Approach

The direct approach we adopt here is the following. Let

$$T_k = \left\{ y \in T : \sum_j a_{kj}/y_j = \max_i \sum_j a_{ij}/y_j \right\} \quad (k = 1, 2, \ldots, s).$$

Then

$$\min_{y \in T_k} \max_i \sum_j a_{ij}/y_j = \min_{y \in T_k} \sum_j a_{kj}/y_j,$$

and we find for each k a $y_k^* \in T_k$ which gives this minimum. This is not altogether easy since usually the minimum in the unrestricted region T does not occur in T_k and so, although the convexity of $\sum a_{kj}/y_j$ in y then allows us to restrict attention to the boundary of T_k, the detection of y_k^* often requires some elaborate enumeration of parts of this boundary. Since $T = \cup T_k$ we can then take y^* as a y_k^* corresponding to the least of

$$\min_{y \in T_k} \sum_j a_{kj}/y_j.$$

We illustrate this method by the following example.

Example 4. Let $t = 6$ and suppose that, although we are interested in all contrasts, our main concern is to detect whether one or other of the inequalities $|\theta_1 - \theta_2| \geqslant \Delta$, $|\theta_1 - \theta_3| \geqslant \Delta$, $|\theta_2 - \theta_3| \geqslant \Delta$, $|\theta_3 - \theta_4| \geqslant \Delta$, $|\theta_3 - \theta_5| \geqslant \Delta$, $|\theta_3 - \theta_6| \geqslant \Delta$ holds.

Here $\mathbf{A} = [a_{ij}]$ is given by

$$\Delta^2 \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

In any example we exploit any symmetries that we recognize, and here it is clear that we can restrict attention to the subset T' of T for which $y_1 = y_2, y_4 = y_5 = y_6$, so that the problem is reduced to that of determining

$$\min_{y \in T'} \max (2/y_1, 1/y_1 + 1/y_3, 1/y_3 + 1/y_4).$$

Let $T_1 = \{y \in T' : y_1 \leqslant y_3, 2/y_1 \geqslant 1/y_3 + 1/y_4\}$. Since the minimum over T' of $2/y_1$ occurs outside T_1, we must seek for the minimum over T_1 of $2/y_1$ on the boundary of T_1.

Now this minimum on the boundary of T_1 where $y_1 = y_3$ can be shown to be 12, occurring where $y_1 = y_3 = y_4 = \frac{1}{6}$; also the minimum on the boundary of T_1 where $2/y_1 = 1/y_3 + 1/y_4$ can be shown to be $8+2\sqrt{3}$, occurring where

$$(6+8\sqrt{3})(y_1, y_3, y_4) = (2\sqrt{3}, 3+\sqrt{3}, 1+\sqrt{3}). \tag{3.1}$$

Hence

$$\min_{T_1} \frac{2}{y_1} = 8+2\sqrt{3}$$

with $\mathbf{y}_1^*$ determined by (3.1).

After similarly tedious calculation we can show that

$$\min_{T_1}\left(\frac{1}{y_1}+\frac{1}{y_3}\right) = 12, \quad \text{and} \quad \min_{T_1}\left(\frac{1}{y_3}+\frac{1}{y_4}\right) = 8+2\sqrt{3},$$

the latter again occurring at (3.1). Hence the optimal design is given by (3.1) with min–max $(8+2\sqrt{3})/\Delta^2$.

4. The Game Associated with the Design Problem

4.1. *General Remarks*

We now show that the design problem (2.2) is greatly simplified when we recognize its equivalence to that of finding an optimal strategy in a certain finite, but non-matrix, game. We first list together for easy reference some points of notation and definitions of various sets and functions which enter our argument. We also precede our mathematical results with an explanation of their motivation and their practical consequences; the reader who has then no appetite for the mathematical proofs that follow may proceed immediately to the computational routine of section 5.

Notation. We use the vector notation $\mathbf{v} > \mathbf{0}$ to mean that each component $v_i \geqslant 0$ and at least one $v_i > 0$, and $\mathbf{v} \gg \mathbf{0}$ to mean that each component $v_i > 0$. The symbols Σ_i and Σ_j denote summation over $i = 1, 2, \ldots, s$ and $j = 1, 2, \ldots, t$ respectively. To save tiresome repetition we use max and min, without specifying the sets involved, to denote maximum over S and minimum over T, respectively, where S and T are the two simplexes defined by (4.1) and (4.2) below; otherwise the sets will be specified.

The matrix A. The $s \times t$ matrix $\mathbf{A} = [a_{ij}]$ of non-negative elements is such that for each i, $a_{ij} > 0$ for some i, and for each j, $a_{ij} > 0$ for some j.

Sets. We require the two simplexes in R^s and R^t defined by

$$S = \{\mathbf{x} \in R^s : \mathbf{x} > \mathbf{0}, \quad \sum_i x_i = 1\}, \tag{4.1}$$

$$T = \{\mathbf{y} \in R^t : \mathbf{y} > \mathbf{0}, \quad \sum_j y_j = 1\}. \tag{4.2}$$

Functions. The interrelations between three real-valued functions K, L, M and a vector-valued function $\mathbf{Y}$ play a central role in the theory. The following are the defining relations:

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i,j} a_{ij} x_i / y_j \quad (\mathbf{x} \in S, \mathbf{y} \in T), \tag{4.3}$$

with the interpretation that $a_{ij} x_i / y_j = 0$ when either $a_{ij} = 0$ or $x_i = 0$, even if $y_j = 0$. Note that $K(\mathbf{x}, \mathbf{y})$ can assume the value ∞ (for example, when $y_j = 0$ and $a_{ij} > 0, x_i > 0$).

Useful elementary reformulations of this definition are

$$K(\mathbf{x},\mathbf{y}) = \sum_i x_i a_i(\mathbf{y}) \quad (\mathbf{x}\in S,\ \mathbf{y}\in T), \tag{4.4}$$

where $a_i(\mathbf{y}) = \Sigma_j a_{ij}/y_j$ and with the interpretation $x_i a_i(\mathbf{y}) = 0$ when $x_i = 0$, even if $a_i(\mathbf{y}) = \infty$; and

$$K(\mathbf{x},\mathbf{y}) = \sum_j \frac{1}{y_j} \sum_i a_{ij} x_i \quad (\mathbf{x}\in S,\ \mathbf{y}\in T), \tag{4.5}$$

with the convention that

$$\left(\sum_i a_{ij} x_i\right) \Big/ y_j = 0 \quad \text{if} \quad \sum_i a_{ij} x_i = 0.$$

Further

$$L(\mathbf{x}) = \sum_j \left(\sum_i a_{ij} x_i\right)^{\frac{1}{2}} \quad (\mathbf{x}\in S), \tag{4.6}$$

$$M(\mathbf{x}) = \{L(\mathbf{x})\}^2 \quad (\mathbf{x}\in S). \tag{4.7}$$

Since $\mathbf{A}'\mathbf{x} > \mathbf{0}$ for every $\mathbf{x}\in S$, it follows that $L(\mathbf{x}) > 0$ for every $\mathbf{x}\in S$, and so we can define

$$Y_j(\mathbf{x}) = \left(\sum_i a_{ij} x_i\right)^{\frac{1}{2}} \Big/ L(\mathbf{x}) \quad (\mathbf{x}\in S), \tag{4.8}$$

and hence

$$\mathbf{Y}(\mathbf{x}) = [Y_1(\mathbf{x})\, Y_2(\mathbf{x}) \ldots Y_t(\mathbf{x})]' \quad (\mathbf{x}\in S). \tag{4.9}$$

4.2. *Motivation and Summary of Results*

In the design problem (2.2) we are already faced with a problem in game theory. For $\mathbf{y}^*$ is a min–max strategy for Player II in the game $(S_0, T, a_i(\mathbf{y}))$, where

(i) Player I has a space S_0 of s pure strategies, for convenience denoted by the s vertices of S, namely $\mathbf{e}_1 = [1\,0 \ldots 0]', \ldots, \mathbf{e}_s = [0\,0 \ldots 1]'$;

(ii) Player II has mixed strategy space T;

(iii) the payoff (the amount paid by II to I) is $a_i(\mathbf{y})$ when I chooses $\mathbf{e}_i$ and II chooses $\mathbf{y}$.

The inherent difficulty of the problem is now more evident. It lies in the awkwardness of handling the discrete space S_0 and in the related fact that, in general,

$$\max_i \min a_i(\mathbf{y}) < \min \max_i a_i(\mathbf{y})$$

so that the game is not determined. For instance, in Example 3,

$$\max_i \min a_i(\mathbf{y}) = 4/\Delta^2$$

by (2.3) and

$$\min \max_i a_i(\mathbf{y}) = (8+2\sqrt{3})/\Delta^2$$

as found in section 3. Hence for $(S_0, T, a_i(\mathbf{y}))$ there are no optimal (as distinct from max–min and min–max) strategies for the two players and so there is no advantage to be gained by studying the game from the viewpoint of Player I at this stage.

Our first task is thus to follow through the consequences of removing this discrete feature by the device of replacing S_0 by the corresponding mixed strategy space S, with strategies

$$\mathbf{x} = \sum_i x_i \mathbf{e}_i \quad \left(x_i \geqslant 0, \sum_i x_i = 1\right),$$

and hence the payoff $a_i(\mathbf{y})$ by $K(\mathbf{x}, \mathbf{y})$. The effect of this is happily that Player II's min–max behaviour is unaltered (Theorem 2). Any direct attack on this problem will, of course, lead us back to the approach of section 3; our one hope of a simplification is that now we may discover Player II's min–max strategy by studying Player I's max–min problem.

We must therefore now pose the following three questions:

(i) Is the new game (S, T, K) determined; is $\min \max K(\mathbf{x}, \mathbf{y}) = \max \min K(\mathbf{x}, \mathbf{y})$?

(ii) If so, can optimal x-strategies be readily found?

(iii) If so, can an optimal y-strategy be easily constructed from an optimal x-strategy; and is this strategy unique?

The three parts of Theorem 3 answer these questions in the affirmative. Part (i) depends essentially on the convexity of the functions $a_i(\mathbf{y})$, which in turn provide the convex–concave property of $K(\mathbf{x}, \mathbf{y})$, as stated in Lemma 3a. The answer to (ii) provides the basis of the computational scheme of section 5. The set of x-optimal strategies is found to be those $\mathbf{x}^* \in S$ at which $L(\mathbf{x})$ attains its maximum over S. Parts (ii) and (iii) depend on the inter-relations of the functions $K, \mathbf{Y}$ and M (and hence L) and this is the motivation for establishing Lemma 3b before Theorem 3. The outcome of (iii) is that the unique y-strategy $\mathbf{y}^*$ is given by $\mathbf{y}^* = Y(\mathbf{x}^*)$, where $\mathbf{x}^*$ is any optimal x-strategy. Although uniqueness is not an essential feature it is useful to know that the design problem is completely solved when we find any one $\mathbf{x}^*$.

The advantage of the game technique is that the computational problem is reduced to that of maximizing the concave function $L(\mathbf{x})$ over the convex set S, for which a straightforward technique can be devised. As opposed to the direct approach we are here dealing with the single simple function $L(\mathbf{x})$ instead of the much more complicated function, the maximum over i of $a_i(\mathbf{y})$. The fact that often $s < t$ or that there are symmetries in the problem may make the game approach even more favourable. To sum up, it is our ability to answer questions (i), (ii) and at least the first part of (iii) in the affirmative that has enabled us to simplify the computational problem. This is largely due to the exploitation of the convexity and special form of the $a_i(\mathbf{y})$, which enable us to obtain the specific forms of $\mathbf{Y}(\mathbf{x})$ and $M(\mathbf{x})$ in Lemma 3b; the use of the technique of this paper for other design problems would depend on some similar form of exploitation.

4.3. *Mathematical Results*

Theorem 2. If $\mathbf{y}^* \in T$ satisfies one of the relations

(a) $\max_i a_i(\mathbf{y}^*) = \min \max_i a_i(\mathbf{y})$,

(b) $\max K(\mathbf{x}, \mathbf{y}^*) = \min \max K(\mathbf{x}, \mathbf{y})$,

then (i) it satisfies the other, (ii) $\mathbf{y}^* \geqslant \mathbf{0}$.

Proof. (i) This is established if we can show that

$$\max_{\mathbf{x}} K(\mathbf{x}, \mathbf{y}) = \max_i a_i(\mathbf{y})$$

for every $\mathbf{y} \in T$, for then, in particular,

$$\max_{\mathbf{x}} K(\mathbf{x}, \mathbf{y}^*) = \max_i a_i(\mathbf{y}^*),$$

and also

$$\min_{\mathbf{y}} \max_{\mathbf{x}} K(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{y}} \max_i a_i(\mathbf{y}).$$

Suppose that for any given $\mathbf{y} \in T$,

$$\max_i a_i(\mathbf{y}) = a_k(\mathbf{y});$$

then for every $\mathbf{x} \in S$, $K(\mathbf{x}, \mathbf{y}) \leqslant a_k(\mathbf{y})$ since

$$\begin{aligned} a_k(\mathbf{y}) - K(\mathbf{x}, \mathbf{y}) &= a_k(\mathbf{y}) \sum_i x_i - \sum_i x_i a_i(\mathbf{y}) \\ &= \sum_i x_i \{a_k(\mathbf{y}) - a_i(\mathbf{y})\} \geqslant 0. \end{aligned}$$

Also $K(\mathbf{x}, \mathbf{y}) = a_k(\mathbf{y})$ when $x_k = 1, x_i = 0$ $(i \neq k)$. Hence the result.

(ii) Our result here follows at once if we can show that $y_j^* = 0$ for some j implies that

$$\max_i a_i(\mathbf{y}^*) \neq \min_{\mathbf{y}} \max_i a_i(\mathbf{y}).$$

Now if $y_j^* = 0$ then $a_i(\mathbf{y}^*) = \infty$ for some i, since $a_{ij} > 0$ for some i. Hence

$$\max_i a_i(\mathbf{y}^*) = \infty.$$

But, by (2.3),

$$\min_{\mathbf{y}} \max_i a_i(\mathbf{y}) \leqslant t \max_i \sum_j a_{ij} < \infty.$$

Lemma 3a. (i) On S, $K(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{x}$ for each given $\mathbf{y} \in T$.

(ii) On T, $K(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{y}$ for each given $\mathbf{x} \in S$.

Proof. (i) By definition (4.4), $K(\mathbf{x}, \mathbf{y})$ is clearly linear in $\mathbf{x}$ for given $\mathbf{y} \in T$, that is

$$K\{\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \mathbf{y}\} = \lambda K(\mathbf{x}_1, \mathbf{y}) + (1-\lambda) K(\mathbf{x}_2, \mathbf{y}),$$

and trivially

$$K\{\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \mathbf{y}\} \geqslant \lambda K(\mathbf{x}_1, \mathbf{y}) + (1-\lambda) K(\mathbf{x}_2, \mathbf{y}) \quad (0 < \lambda < 1).$$

(ii) The convexity of $K(\mathbf{x}, \mathbf{y})$ in $\mathbf{y}$ for given $\mathbf{x} \in S$ follows at once from the definition (4.5) and the inequality

$$\frac{1}{\lambda y_{1j} + (1-\lambda) y_{2j}} \leqslant \frac{\lambda}{y_{1j}} + \frac{(1-\lambda)}{y_{2j}} \quad (0 < \lambda < 1).$$

Lemma 3b. For every $\mathbf{x} \in S$ there corresponds a unique $\mathbf{y} = \mathbf{Y}(\mathbf{x}) \in T$ such that

$$K\{\mathbf{x}, \mathbf{Y}(\mathbf{x})\} = \min_{\mathbf{y}} K(\mathbf{x}, \mathbf{y}) = M(\mathbf{x}).$$

Proof. For given $\mathbf{x} \in S$, $K(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{y} \in T$. Clearly from definition (4.5), in order to minimize $K(\mathbf{x}, \mathbf{y})$ we must put $y_j = 0$ if

$$\sum_i a_{ij} x_i = 0.$$

If we then minimize

$$\sum_{j \in J(\mathbf{x})} \frac{1}{y_j} \sum_i a_{ij} x_i,$$

where

$$J(\mathbf{x}) = \left\{ j : \sum_i a_{ij} x_i > 0 \right\},$$

subject to

$$\sum_{j \in J(\mathbf{x})} y_j = 1,$$

by a direct application of the Lagrange-multiplier technique, we find that there is one local (hence the absolute) minimum, occurring where

$$y_j \propto \left(\sum_i a_{ij} x_i \right)^{\frac{1}{2}} \quad \{ j \in J(\mathbf{x}) \}.$$

Thus, since $Y_j(\mathbf{x}) = 0$ if $j \notin J(\mathbf{x})$, the minimum of $K(\mathbf{x}, \mathbf{y})$ over T occurs at the unique point $\mathbf{y} = \mathbf{Y}(\mathbf{x})$, and we find by substitution that $K(\mathbf{x}, \mathbf{Y}(\mathbf{x})) = M(\mathbf{x})$.

Theorem 3. For the game (S, T, K)

(i) $\min \max K(\mathbf{x}, \mathbf{y}) = \max \min K(\mathbf{x}, \mathbf{y})$;

(ii) the set of optimal x-strategies is $\{\mathbf{x}^* \in S : L(\mathbf{x}^*) = \max L(\mathbf{x})\}$;

(iii) there is a unique optimal y-strategy $\mathbf{y}^*$ given by $\mathbf{y}^* = \mathbf{Y}(\mathbf{x}^*) \geqslant \mathbf{0}$, where $\mathbf{x}^*$ is any optimal x-strategy.

Proof. (i) To establish this result we can appeal to an unpublished proof due to M. Schiffman of a general min–max theorem; we refer here to the version given in the second proof of Theorem 1.5.1 of Karlin (1959, pp. 28–30). The first part of that proof yields the following result for a real-valued function $K(\mathbf{x}, \mathbf{y})$, defined for $\mathbf{x} \in S, \mathbf{y} \in T$. If

(a) S, T are closed, bounded, convex sets;

(b) $K(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{y}$ for each $\mathbf{x}$, and concave in $\mathbf{x}$ for each $\mathbf{y}$;

(c) $K(\mathbf{x}, \mathbf{y})$ is continuous and the convexity and concavity of $K(\mathbf{x}, \mathbf{y})$ are strict, then $\min \max K(\mathbf{x}, \mathbf{y}) = \max \min K(\mathbf{x}, \mathbf{y})$.

Our particular S, T, K satisfy (a) and (b), but not (c). In the general case, however, property (c) is used only to establish that, for each $\mathbf{x} \in S$, there corresponds a unique $\mathbf{Y}(\mathbf{x}) \in T$ such that $K(\mathbf{x}, \mathbf{Y}(\mathbf{x})) = \min K(\mathbf{x}, \mathbf{y})$, and that $\mathbf{Y}(\mathbf{x})$ and $M(\mathbf{x}) = \min K(\mathbf{x}, \mathbf{y})$ are continuous. Lemma 3b thus replaces this part of the general proof, for our $\mathbf{Y}(\mathbf{x})$ and $M(\mathbf{x})$ are clearly continuous. Thus our knowledge of the special form of K allows us to dispense with condition (c), and the result follows.

(ii) $\mathbf{x}^*$ is an optimal x-strategy if and only if $\min K(\mathbf{x}^*, \mathbf{y}) = \max \min K(\mathbf{x}, \mathbf{y})$, i.e. by Lemma 3b, if and only if $M(\mathbf{x}^*) = \max M(\mathbf{x})$ or $L(\mathbf{x}^*) = \max L(\mathbf{x})$.

(iii) Let $\mathbf{x}^*$ be any fixed optimal x-strategy. If $\mathbf{y}^* \in T$ is any optimal y-strategy then $K(\mathbf{x}^*, \mathbf{y}^*) \leqslant K(\mathbf{x}^*, \mathbf{y})$ for every $\mathbf{y} \in T$ and so $K(\mathbf{x}^*, \mathbf{y}^*) = \min K(\mathbf{x}^*, \mathbf{y})$. By Lemma 3b the only solution for $\mathbf{y}^*$ is $Y(\mathbf{x}^*)$, and the uniqueness property is proved. (Note that we have incidentally established the interesting property of L and Y that $L(\mathbf{x}_1) = L(\mathbf{x}_2) = \max L(\mathbf{x})$ implies $Y(\mathbf{x}_1) = Y(\mathbf{x}_2)$.)

The optimal solution $\mathbf{y}^*$ satisfies $\max K(\mathbf{x}, \mathbf{y}^*) = \min \max K(\mathbf{x}, \mathbf{y})$, and so, by Theorem 2, $\mathbf{y}^* \geqslant \mathbf{0}$.

Combining Theorems 2 and 3, we then have the following corollary, which forms the basis of the computational procedure of the next section.

Corollary. The unique solution $\mathbf{y}^*$ of the min–max design problem is given by $\mathbf{y}^* = Y(\mathbf{x}^*)$, where $\mathbf{x}^*$ is any solution of $L(\mathbf{x}^*) = \max L(\mathbf{x})$, and the min–max is $\min \max K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x}^*, \mathbf{y}^*) = M(\mathbf{x}^*)$.

5. Computational Scheme

Our remaining task is to devise a method of calculating an $\mathbf{x}^* \in S$ such that $L(\mathbf{x}^*) = \max L(\mathbf{x})$. Since the optimal y-strategy $\mathbf{y}^* = \mathrm{Y}(\mathbf{x}^*) \geqslant \mathbf{0}$ we know that $\mathrm{A}'\mathbf{x}^* \geqslant \mathbf{0}$, and so, when convenient, we can confine the $\mathbf{x}$ we consider to the open convex set $\{\mathbf{x} \in S : \mathrm{A}'\mathbf{x} \geqslant \mathbf{0}\} \subset S$.

Our simple search method for locating an $\mathbf{x}^*$ uses the equivalence of the two properties stated in Theorem 4. The proof of this theorem is of little direct interest to us and so is omitted. It depends on either an appeal to standard results in concave programming (see, for example, Karlin (1959), pp. 199–204) or a straightforward application of the concavity of $L(\mathbf{x})$ and the property $\Sigma_i x_i L^{(i)}(\mathbf{x}) = \frac{1}{2} L(\mathbf{x})$, where the superscript notation denotes partial differentiation with respect to x_i.

Theorem 4. The following two statements are equivalent.

(i) $\mathbf{x}^* \in S$ and $L(\mathbf{x}^*) = \max L(\mathbf{x})$;

(ii) $\mathbf{x}^* \in S$ and λ^* can be found to satisfy

$$L^{(i)}(\mathbf{x}^*) - \lambda^* \leqslant 0 \quad (i = 1, 2, \ldots, s), \tag{5.1}$$

$$\tfrac{1}{2} L(\mathbf{x}^*) - \lambda^* = 0. \tag{5.2}$$

Moreover, if in (ii) $L^{(i)}(\mathbf{x}^*) - \lambda^* < 0$, then $x_i^* = 0$.

We now present the complete computational scheme and then discuss its validity.

Computational Scheme. Denote by $L_{kl\ldots}(\mathbf{x})$ the expression obtained from $L(\mathbf{x})$ by setting $x_k = x_l = \ldots = 0$.

Step 1. By actually solving, determine whether the equations

$$L^{(i)}(\mathbf{x}) - \lambda = 0 \quad (i = 1, 2, \ldots, s), \tag{5.3}$$

$$\sum_i x_i = 1 \tag{5.4}$$

have a solution $(\mathbf{x}_0, \lambda_0)$ with $\mathbf{x}_0 > \mathbf{0}$. If so, then take $\mathbf{x}^* = x_0$; otherwise proceed to Step 2.

Step 2. Carry out the following investigation for each k.

Determine, by actually solving, whether the equations

$$L_k^{(i)}(\mathbf{x}) - \lambda = 0 \quad (i \neq k), \tag{5.5}$$

$$\sum_{i \neq k} x_i = 1, \tag{5.6}$$

$$x_k = 0 \tag{5.7}$$

have a solution $(\mathbf{x}_k, \lambda_k)$ with $\mathbf{x}_k > \mathbf{0}$.

For those k for which $\mathbf{x}_k$ exists, calculate the corresponding $L_k(\mathbf{x}_k) = 2\lambda_k$ and proceed to Step 3.

Step 3. Carry out the following investigation for each (k,l) for which $\mathbf{x}_k$ and $\mathbf{x}_l$ do not exist.

Determine, by actually solving, whether the equations

$$L_{kl}^{(i)}(\mathbf{x}) - \lambda = 0 \quad (i \neq k, l), \tag{5.8}$$

$$\sum_{i \neq k,l} x_i = 1, \tag{5.9}$$

$$x_k = x_l = 0 \tag{5.10}$$

have a solution $(\mathbf{x}_{kl}, \lambda_{kl})$ with $\mathbf{x}_{kl} > \mathbf{0}$.

For those (k,l) for which $\mathbf{x}_{kl}$ exists, calculate the corresponding $L_{kl}(\mathbf{x}) = 2\lambda_{kl}$. Then proceed to the next obvious step of investigating the $L_{klm}(\mathbf{x})$ for which there are no $\mathbf{x}_k$, $\mathbf{x}_l$, $\mathbf{x}_m$, $\mathbf{x}_{kl}$, $\mathbf{x}_{km}$, or $\mathbf{x}_{lm}$.

Continue this finite process until it terminates. Scan the calculated set of $L_k(\mathbf{x}_k)$'s, $L_{kl}(\mathbf{x}_{kl})$'s, etc., and choose as $\mathbf{x}^*$ an $\mathbf{x}_k, \mathbf{x}_{kl}, \ldots$, which corresponds to the maximum of this set.

Note: If $x_k = x_l = \ldots = 0$ implies that a component of $\mathbf{A}'\mathbf{x}$ is zero then by the remark at the beginning of this section, the corresponding $L_{kl\ldots}(\mathbf{x})$ need not be investigated.

Validity of the computational scheme. This routine, which is usually easy to apply since it depends only on solving elementary sets of equations, can be justified as follows.

If, in Step 1, $\mathbf{x}_0 > \mathbf{0}$ exists then $\mathbf{x}_0 \in S$ by (5.4). Also, trivially from (5.3), $L^{(i)}(\mathbf{x}_0) - \lambda_0 \leqslant 0$ $(i = 1, 2, \ldots, s)$ and we have $\frac{1}{2}L(\mathbf{x}_0) = \Sigma_i x_{0i} L^{(i)}(\mathbf{x}_0) = \lambda_0 \Sigma_i x_{0i} = \lambda_0$. Hence $\mathbf{x}^* = \mathbf{x}_0$ has property (ii) of Theorem 4, and so has also property (i).

If no $\mathbf{x}_0$ exists it follows from Theorem 4 that $L^{(k)}(\mathbf{x}^*) - \lambda^* < 0$ for some k, and thus that $x_k^* = 0$ for some k. Hence we may confine attention in Step 2 to the situation on each of the boundaries $S_k = \{\mathbf{x} \in S : x_k = 0\}$. On S_k we can replace $L(\mathbf{x})$ by $L_k(\mathbf{x})$. Theorem 4 then applies to $L_k(\mathbf{x})$ with S replaced by S_k and $(i = 1, 2, \ldots, s)$ by $(i \neq k)$. Thus, in Step 2, by reasoning similar to that applied in Step 1, the $\mathbf{x}_k > \mathbf{0}$ we obtain are such that

$$L(\mathbf{x}_k) = L_k(\mathbf{x}_k) = \max_{\mathbf{x} \in S_k} L_k(\mathbf{x}) = \max_{\mathbf{x} \in S_k} L(\mathbf{x}).$$

For any k for which $\mathbf{x}_k > 0$ does not exist

$$\max_{\mathbf{x} \in S_k} L_k(\mathbf{x})$$

must occur on one of the subsets $S_{kl} = \{\mathbf{x} \in S_k : x_l = 0\}$, and clearly we need consider only those l for which $\mathbf{x}_l > 0$ also does not exist. The $\mathbf{x}_{kl}$ we obtain are such that

$$L(\mathbf{x}_{kl}) = \max_{\mathbf{x} \in S_{kl}} L(\mathbf{x}).$$

So, by continuing the process and choosing $\mathbf{x}^*$ as we do, we arrive at our optimal $\mathbf{x}$-strategy.

We can see heuristically that $x_k^* = 0$ only when the corresponding effect is unimportant compared with the other effects. In many practical situations there will be few unimportant specified effects and so we may expect the process to terminate very quickly.

6. Applications of the Game Approach

The technique in its various aspects is adequately demonstrated by its application to the following simple examples.

Example 1. (See section 1.) By symmetry we must clearly have

$$x_1^* = x_2^* = \ldots = x_s^* = 1/s$$

and so

$$s\Delta^2 \mathbf{A}'\mathbf{x}^* = [s\,1\,1 \ldots 1]',$$

giving

$$y_1^* = \sqrt{(s)}/\{s+\sqrt{(s)}\}, \quad y_2^* = y_3^* = \ldots = y_t^* = 1/\{s+\sqrt{(s)}\},$$

the well-known optimal design with min–max $\{1+\sqrt{(s)}\}^2/\Delta^2$.

Example 4. (See section 3.) Here

$$\Delta L(\mathbf{x}) = (x_1+x_2)^{\frac{1}{2}} + (x_1+x_3)^{\frac{1}{2}} + (x_2+x_3+x_4+x_5+x_6)^{\frac{1}{2}} + x_4^{\frac{1}{2}} + x_5^{\frac{1}{2}} + x_6^{\frac{1}{2}}$$

is symmetric in x_2, x_3 and in x_4, x_5, x_6 and so it is clear that $x_2^* = x_3^*, x_4^* = x_5^* = x_6^*$. In our computations then we can impose the extra condition

$$x_2 = x_3; \quad x_4 = x_5 = x_6. \tag{5.11}$$

Thus, in Step 1, equation (5.4) becomes $x_1 + 2x_2 + 3x_4 = 1$, and in Step 2 we need consider only $L_1(\mathbf{x}), L_2(\mathbf{x}), L_4(\mathbf{x})$.

Step 1. Equations (5.3) and (5.4) reduce to

$$\begin{aligned}
(x_1+x_2)^{-\frac{1}{2}} &\qquad\qquad -\lambda\Delta = 0,\\
\tfrac{1}{2}(x_1+x_2)^{-\frac{1}{2}} + \tfrac{1}{2}(2x_2+3x_4)^{-\frac{1}{2}} &\qquad -\lambda\Delta = 0,\\
\tfrac{1}{2}(2x_2+3x_4)^{-\frac{1}{2}} + \tfrac{1}{2}x_4^{-\frac{1}{2}} &- \lambda\Delta = 0,\\
x_1 + 2x_2 + 3x_4 &\qquad = 1,
\end{aligned}$$

and have unique solution $(x_1, x_2, x_4) = (\frac{2}{3}, -\frac{1}{3}, \frac{1}{3}) \not> 0$; hence we proceed to Step 2.

Step 2. Here by the Note there is no need to consider $L_4(\mathbf{x})$.

(i) For $k = 1$ equations (5.5)–(5.7) reduce to

$$\begin{aligned} \tfrac{1}{2}x_2^{-\frac{1}{2}} + \tfrac{1}{2}(2x_2+3x_4)^{-\frac{1}{2}} - \lambda\Delta &= 0, \\ \tfrac{1}{2}(2x_2+3x_4)^{-\frac{1}{2}} + \tfrac{1}{2}x_4^{-\frac{1}{2}} - \lambda\Delta &= 0, \\ 2x_2+3x_4 &= 1, \\ x_1 &= 0, \end{aligned}$$

which have unique solution $(x_1, x_2, x_4) = (0, \frac{1}{5}, \frac{1}{5}) > \mathbf{0}$ and we have

$$L_1(\mathbf{x}_1) = 2\lambda_1 = \{1+\sqrt{5}\}/\Delta.$$

(ii) For $k = 2$ equations (5.5)–(5.7) yield

$$\begin{aligned} x_1^{-\frac{1}{2}} - \lambda\Delta &= 0, \\ \tfrac{1}{2}(1+\tfrac{1}{3}\sqrt{3})x_4^{-\frac{1}{2}} - \lambda\Delta &= 0, \\ x_1+3x_4 &= 1, \\ x_2 &= 0, \end{aligned}$$

with unique solution $(x_1, x_2, x_4) = (6, 0, 2+\sqrt{3})/\{3(4+\sqrt{3})\} > \mathbf{0}$ and we have

$$L_2(\mathbf{x}_2) = 2\lambda_2 = (8+2\sqrt{3})^{\frac{1}{2}}/\Delta.$$

Here for each possible k $(= 1, 2)$ we have found an $\mathbf{x}_k > \mathbf{0}$ and so we need proceed no further. Since $(8+2\sqrt{3})^{\frac{1}{2}} > 1+\sqrt{5}$ we take $\mathbf{x}^* = \mathbf{x}_2$, and this gives, by the use of $\mathbf{x}^* = \mathbf{Y}(\mathbf{x}^*)$, the optimal design given by (3.1).

Example 5. As an illustration of the use of Step 2 we outline the application of the routine to the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Step 1. Equations (5.3) and (5.4) have no finite solution.

Step 2. By the Note, $L_4(\mathbf{x})$ need not be considered. Only for $k = 2$ do we find an $\mathbf{x}_k > \mathbf{0}$. We have $\mathbf{x}_2 = [0, 0, \frac{1}{2}, \frac{1}{2}]$ with $L_2(\mathbf{x}_2) = 2\sqrt{2}$.

Step 3. By the Note and the existence of $\mathbf{x}_2$ we need consider only $L_{13}(\mathbf{x})$. We easily find $\mathbf{x}_{13} = [0, \frac{4}{5}, 0, \frac{1}{5}]'$ with $L_{13}(\mathbf{x}_{13}) = 1+\sqrt{5}$.

Since $1+\sqrt{5} > 2\sqrt{2}$ we take $\mathbf{x} = \mathbf{x}_{13}$ and so the optimal design is

$$\mathbf{y}^* = [2, \sqrt{5}, 2, 1]'/\{5+\sqrt{(5)}\}.$$

We could in this case have foreseen that $x_1^* = x_3^* = 0$ by observing that the first and third effects are unimportant since $a_1(\mathbf{y}) \leqslant a_2(\mathbf{y}), a_3(\mathbf{y}) \leqslant a_2(\mathbf{y})$ for every $\mathbf{y} \in T$.

ACKNOWLEDGEMENTS

I wish to thank the referee for his constructive criticism of an earlier version of this paper, and also Dr S. D. Silvey for his helpful comments and discussion.

REFERENCES

ELFVING, G. (1959), "Design of linear experiments", in *Probability and Statistics, The Harald Cramér Volume*; editor, U. Grenander. New York: Wiley.

GUEST, P. G. (1958), "The spacing of observations in polynomial regression", *Ann. math. Statist.*, **29**, 294–299.

HOEL, P. G. (1958), "Efficiency problems in polynomial estimation", *Ann. math. Statist.*, **29**, 1134–1146.

KARLIN, S. (1959), *Mathematical Methods and Theory in Games, Programming, and Economics*. London: Pergamon.

KIEFER, J. (1958), "On the nonrandomized optimality and randomized nonoptimality of symmetric designs", *Ann. math. Statist.*, **29**, 675–699.

—— and WOLFOWITZ, J. (1959), "Optimum designs in regression problems", *Ann. math. Statist.*, **30**, 271–294.

SCHEFFÉ, H. (1953), "A method for judging all contrasts in the analysis of variance", *Biometrika*, **40**, 87–104.

WILLIAMS, E. J. (1958), "Optimum allocation for estimation of polynomial regression" (abstract), *Biometrics*, **14**, 573.

6 PARAMETRIC TOLERANCE REGIONS

6.1 Background

In the 1960's the Hospital Engineering Research Unit, later the Building Services Research Unit, of the University of Glasgow, was involved in problems of the design of an adequate supply of engineering services, such as gas, electricity, water, piped oxygen, to meet demand in new hospitals. Data were available on patterns of demand in situations where there was deliberate over-supply. Some of the problems seemed to call for the standard technique of guaranteed-cover statistical tolerance regions. The rationale underlying these was carefully explained many times and in many different ways and sets of tables provided to allow easy estimation of design values.

In terms of the general setting of parametric statistical problems in §1 the experiment f refers to the future demand situation and the experiment e to the observed demand pattern data. A guaranteed-coverage tolerance region $R(x) \subset Y$ is then sought which satisfies the probabilistic statement

$$P_e[x : P_f\{R(x)|\theta\} \geqq c|\theta] \geqq g \text{ for all } \theta \in \Theta,$$

where suffices e and f are used to make clear the nature of the probability measures involved. Thus the guarantee g is used in the usual frequentist confidence coefficient sense, referring to repetitions of e, and c is the coverage or proportion of future 'demands' which are to be met at this confidence level. The difficulty of interpretation of this approach, particularly in

relation to g, has been argued at length in Aitchison (7:1964). The principal confusion is that the probabilistic interpretation of g, which is strictly in terms of repetitions of the informative experiment e, is often transferred by users, such as the hospital engineers, to the proportion of successes that would be obtained in the use of a specific design: 'If we use R to set designs for 100 hospitals then, roughly speaking, 100g of these hospitals will function satisfactorily as regards coverage.'

6.2 A first step towards a solution

Faced with persistent misinterpretations such as this, the consulting statistician must eventually face the self-criticising question: Have I really formulated this problem in an appropriate way? In a rethinking of such problems Aitchison (7:1964) makes the following points.

1. Such design problems are essentially decision problems and so there may be considerable advantages, not least in sensible communication between client and statistician, in formulating them in proper decision theoretic terms.
2. Since the action space is the space of sets in Y the utility function takes the form $U(R,\theta)$, where $R \subset Y$. In more practically oriented terms it is, however, easier to adopt a utility structure of the form $V(R,y)$, denoting the advantage of the design region R when the actual outcome in the future experiment f is y. This confrontation between R and y is probably the easiest for the designer to appreciate and V can be easily related to U by

$$U(R,\theta) = \int_Y V(R,y)p(y|\theta)\,dy.$$

3. The realisation that decision theory is much more easily and sensibly pursued from a Bayesian viewpoint suggests the adoption of this approach, thus choosing a design region R to maximise

$$W(R,x) = \int_{\Theta} U(R,\theta)p(\theta|x)d\theta$$

$$= \int_{\Theta} \{\int_{Y} V(R,y)p(y|\theta)dy\}p(\theta|x)d\theta$$

$$= \int_{Y} V(R,y)\{\int_{\Theta} p(y|\theta)p(\theta|x)d\theta\}dy$$

$$= \int V(R,y)p(y|x)dy,$$

where

$$p(y|x) = \int_{\Theta} p(y|\theta)p(\theta|x)d\theta.$$

It is regrettable, though perhaps inevitable in the light of the prevailing climate in 1964 of confrontation between 'Bayesians' and 'frequentists', that this decision theory approach was not immediately pursued. But fashion dictated that Bayesians should show that they could produce goods similar to frequentists but with, of course, a better brand label. Thus Aitchison (7:1964) did not immediately develop the decision theory approach but diverted to what he called restricted tolerance region problems with utility function

$$U(R,\theta) = \begin{cases} 1 & \text{if } P(R|\theta) \geq c \\ 0 & \text{if } P(R|\theta) < c. \end{cases}$$

Since $W(R,x)$ is maximised trivially and uselessly by $R = Y$ the problem was regarded as finding R, for a specified g, to satisfy

$W(R,x) = g < 1$, the maximum attained when $R = Y$. Leaving aside the argument as to whether the Bayesian presentation is more appropriate than the frequentist this restricted Bayesian tolerance region still suffers from the (c,g) problem, the difficulty of interpreting in real terms statements involving probabilities about probabilities. The difficulty is exposed by the posing of the basic question: is it better to be 95 per cent certain that 99 per cent of future demands will be met or 99 per cent certain that 95 per cent of future demands will be met?

Thus although the technical problems of defining Bayesian tolerance regions and of deriving them for standard situations had been resolved it could be argued that these regions provided no better a resolution of the real problems than classical statistical tolerance regions. But in this study a more practical and sensible approach had been touched on, through the use of statistical decision theory. It is towards the problems raised by this approach that we turn our attention in the next section.

AITCHISON, J. (1964)

Bayesian tolerance regions (with discussion)

Reprinted from *J. R. Statist. Soc.* B26, 161-72 and 192-210

Two Papers on the Comparison of Bayesian and Frequentist Approaches to Statistical Problems of Prediction

[Read at a RESEARCH METHODS MEETING of the SOCIETY, February 5th, 1964, Professor D. V. LINDLEY in the Chair]

Bayesian Tolerance Regions

By J. AITCHISON

University of Liverpool

SUMMARY

In the theory of statistical tolerance regions, as usually presented in frequentist terms, there are inherent difficulties of formulation, development and interpretation. The present paper re-examines the basic problem from a Bayesian point of view and suggests that such an approach provides a set of widely applicable, mathematically tractable tools, often more tailored to the requirements of users than the corresponding frequentist tools. For the one-dimensional case, Bayesian intervals are quoted for a number of standard distributions and prior densities, and the customary feature of a Bayesian analysis—that special prior densities give rise to standard frequentist results—is briefly demonstrated. A problem which seems to be of greater practical significance, namely the selection of an optimum tolerance region from a set of possible tolerance regions, is also investigated and the overwhelming advantages of the Bayesian approach are indicated.

1. INTRODUCTION

IN this paper the term *Bayesian* refers to any use or user of prior densities on a parameter space, with the associated application of Bayes's theorem, in the analysis of a statistical problem. The term *frequentist* applies to any analysis or analyst of the "objectivist" school, where the use of prior densities is denied and where there is a tendency to interpret probability solely in terms of relative frequencies in large-scale replication. Typical contributions by frequentists on the problems considered here are Wald and Wolfowitz (1946), Bowker (1947), Wallis (1951) and Fraser (1953).

Bayesian revivalists seem so far to have neglected a promising missionary field in the heart of frequentist territory—the field of statistical tolerance regions. Savage (1954, Section 17.3) briefly dismisses the whole concept of tolerance interval as a slippery one, unamenable to behaviouralistic interpretation. Although Raiffa and Schlaifer (1961, Example 6.1.2) introduce their Bayesian treatment of estimation by a simple illustration, which falls into the general class of tolerance-region problems as defined in Section 3 of this paper, their situation does not belong to the special subclass of these problems which is usually understood to constitute the subject matter of tolerance-region theory. Welch and Peers (1963) obtain mainly asymptotic results on conditions under which frequentist and essentially Bayesian upper (or lower) confidence limits for a real parameter θ are identical. Their approach, however, is staunchly frequentist and they repeatedly assert that their weights are not to be regarded as prior probability densities. Since an upper tolerance limit is merely an

upper confidence limit for a quantile of some fixed order it would be possible, at least for cases where quantiles of fixed order are monotonic functions of θ, to derive from their results asymptotic forms of prior density functions which yield frequentist-type limits. While these forms might provide, for such cases, a means of measuring the agreement or disagreement between frequentist and Bayesian analyses they appear to be of limited practical value in the situations treated here.

It is the purpose of this paper to initiate a mildly evangelical campaign among workers in the field of tolerance regions by demonstrating that the Bayesian approach has much to offer in theoretical and practical terms. Since conversion to a new way of thinking often arises out of self-criticism, one of our early tasks is to undertake a critical re-examination of the probabilistic statement which forms the basis of the frequentist structure of tolerance-region theory. This statement is a complicated one, involving many mathematical difficulties and with a frequentist interpretation often misunderstood—in the author's experience—by users of tolerance regions. Bayesian formulations of the problem are mathematically simpler and appear to be much closer to the spirit in which tolerance regions are used in many practical situations. In standard situations the Bayesian, by persuading himself that his prior densities have a special form, may act as if he were a frequentist, though his probabilistic interpretation will differ from the frequentist. We shall note this aspect in our study of standard univariate distributions where one-dimensional intervals are required. When the problem is one of selecting a region from a set of possible tolerance regions on certain optimality criteria—and this, it is suggested, is the type of problem of practical importance—we shall see that Bayesian theory stands up to the task well, whereas it seems difficult even to formulate the problem sensibly in frequentist form.

In defining the experimental setting of the tolerance-region problem we must be careful about notation, since this is crucial to the development and interpretation of the various probabilistic statements we encounter. The notation adopted here is substantially that of Lindley (1961). We suppose that the uncertainty features of a random experiment $\mathscr{E}$, with outcome space $\mathscr{X}$ and event space $\mathscr{B}$, are described by a probability measure on $\mathscr{B}$, say $P(.\,|\,\theta)$ belonging to a parametric family

$$\{P(.\,|\,\theta);\ \theta \in \Theta\},$$

where Θ is the parameter space. We suppose that each measure $P(.\,|\,\theta)$ admits a density function $p(.\,|\,\theta)$ on $\mathscr{X}$. We do not know which particular parameter value θ_0 is the true one. There is available a set $x^{(n)} = (x_1, \ldots, x_n)$ of outcomes of n (independent) replicates of $\mathscr{E}$; we denote the corresponding density on $\mathscr{X}^n$ by $p^{(n)}(x^{(n)}\,|\,\theta)$ and the associated product probability measure by $P^{(n)}(.\,|\,\theta)$. These are the immediate definitions needed by the frequentist. The Bayesian requires the further concept of a prior probability density $\pi(\theta)$ on Θ; this prior density represents either his prior beliefs, if these are ascertainable or a convenient, sensible form of weighting over the possible parameters. The consequence of observing $x^{(n)}$ is to alter, by way of Bayes's theorem, this prior density into a posterior density $\pi(\theta\,|\,x^{(n)})$ over Θ; the correponding probability measure (over the Borel σ-field defined on Θ) we denote by $\Pi(.\,|\,x^{(n)})$. The Bayesian then decides his course of action solely on this posterior distribution and on the consequences of his possible decisions.

We begin our discussion with an examination in Section 2 of the "simple" version of the tolerance-region problem as usually presented by frequentist writers. The problem is to choose some region $R \in \mathscr{B}$, for which it can be reasonably claimed that a proportion c of outcomes of future replicates of $\mathscr{E}$ will fall in R. If θ_0 were known

the tolerance-region user would select a region $R \in \mathscr{B}$ with the property that $P(R \mid \theta_0) = c$. When R has this property we say that R has *cover* c at θ_0. The selection of a suitable R would then usually pose a relatively simple mathematical problem. It is uncertainty about θ which forces some compromise solution on the user. His experimental data $x^{(n)}$ remove some of this uncertainty about θ. The user thus seeks a region $R(x^{(n)}) \in \mathscr{B}$, which he hopes will provide cover at least c; that is, he hopes that he can say with some "confidence" that $P\{R(x^{(n)}) \mid \theta\} \geqslant c$. He would then be reasonably sure that at least a proportion c of future outcomes of $\mathscr{E}$ fall within $R(x^{(n)})$. It is in the probabilistic formulation and interpretation of the term "reasonably sure" that the frequentist and Bayesian begin to differ.

2. The Frequentist Analysis

The basic statement on which the frequentist builds his theory involves the family of probability measures $\{P^{(n)}(. \mid \theta): \theta \in \Theta\}$ associated with the n-replicate outcome space $\mathscr{X}^n$. Let R be a measurable function or statistic, with domain $\mathscr{X}^n$ and range $\mathscr{B}$ and providing, for each $x^{(n)}$, a subset $R(x^{(n)})$ of $\mathscr{X}$. For given θ, cover c and statistic R, denote by $G(\theta, c, R)$ the set of all $x^{(n)}$ which yield, through R, regions of cover at least c at θ, so that

$$G(\theta, c, R) = [x^{(n)} : P\{R(x^{(n)}) \mid \theta\} \geqslant c]. \tag{1}$$

The frequentist says that we have a satisfactory tolerance statistic R if each probability measure of this set is some pre-assigned value q near unity; that is, if

$$P^{(n)}\{G(\theta, c, R) \mid \theta\} = q \tag{2}$$

for *every* $\theta \in \Theta$.

The popular frequentist view of probability—as the counterpart, in a mathematical model of an experiment, of stable relative frequency in a large number of replicates of the experiment—allows the following interpretation. If we repeat the n-replicate experiment a large number of times, each time obtaining some observation $x^{(n)}$ and each time constructing, through R, a tolerance region $R(x^{(n)})$, then a proportion q of these regions will be satisfactory in providing cover at least c; for this interpretation see, for example, Fraser (1957, p. 116) and Weissberg and Beatty (1960). While this statement is comforting to a statistician making repeated use of R to provide his customers with tolerance regions, its meaning for the individual customer is not at all easy to specify. The author has recently experienced the difficulty of trying to sell the frequentist approach to engineers whose work is crucially concerned with problems of tolerance regions. While there was no difficulty in their understanding of the cover c, the confidence coefficient or *quality* q was a much more elusive concept. This is not surprising, for in many applications there is no reality in the idea of repeated replication of the n-replicate experiment. The frequentist might then argue that the probabilistic statement, while not of direct application to the engineer's needs, is intended to give him comfort through the hope that his particular experiment is one of the lucky ones belonging to this proportion q in a population of hypothetical experiments. This seems to be little removed from admitting two different interpretations for probability—the forced acceptance of a kind of degree-of-belief interpretation for $P^{(n)}$ while retaining the relative frequency interpretation for P. It is this confusion of interpretations, none of which quite fits the real problem, which causes misunderstandings with users.

This confusion can be illustrated by a simple example. Suppose that it is required to specify what fixed daily amount of a perishable commodity it is necessary to supply in order to be "reasonably sure" that the probability of supply meeting demand on any one day is c. We may then decide, on the basis of $x^{(n)}$, daily observations (supposed independent) of demand over a month say, to quote a value $r(x^{(n)})$, where $R(x^{(n)}) = \{-\infty, r(x^{(n)})\}$ is a tolerance interval of cover c and quality q. If, for each of 100 months' observations, we construct a supply system with design values based on the function $r(.)$, then the frequentist interpretation is that, roughly speaking, for a proportion q of these systems will the supply be satisfactory in that the probability of supply meeting demand on any day is at least c. Unfortunately, what is sometimes thought to be the interpretation by the user is that, if he observes one month's demands, quotes a supply value from this and then constructs 100 actual systems based on this one supply value, a proportion q of these systems will function satisfactorily as regards cover. This interpretation is clearly wrong; either all are satisfactory or all are not. The user is, however, so confused by the complexity of the probabilistic statement and the insistence on frequentist interpretations that he falls easily into these misunderstandings.

This is, however, not the only difficulty surrounding the statement (2). The probability $P^{(n)}\{G(\theta, c, R) \mid \theta\}$ depends essentially on θ in the general formulation, and the true value θ_0 of the parameter is unknown. Yet the frequentist determines by his probabilistic statement that this probability is to be q, independent of θ. Two escape routes, other than the pursuit of distribution-free regions which reflects a complete change of attitude, are open to avoid this entanglement with θ. He can alter his statement (2) so that the left-hand side does not depend on θ, though his problem of determining a whole function will remain. He can achieve this by requiring R to satisfy

$$\inf_{\theta \in \Theta} P^{(n)}\{G(\theta, c, R) \mid \theta\} = q, \tag{3}$$

as in Fraser (1957, Definition 5.1). This essentially minimax procedure can place undue emphasis on particular values of θ with possibly awkward effects (Lehmann, 1959, p. 13). Alternatively he could introduce a weighting factor $\pi(\theta)$ associated with each θ and choose R so that

$$\int_{\Theta} d\theta\, \pi(\theta) P^{(n)}\{G(\theta, c, R) \mid \theta\} = q, \tag{4}$$

thus producing "Bayes solutions". In either case it will be observed that he is agreeing to place more emphasis on some values of θ than others and so is moving towards a Bayesian outlook. It is not surprising, therefore, that he tends to avoid this escape route and concentrate on the second.

The second approach is to attempt to choose the function R in such a way that $P^{(n)}\{G(\theta, c, R) \mid \theta\}$ is indeed independent of θ. The frequentist thus directs his attention to the search for such a pivotal statistic. For example, in the well-known case where $p(x \mid \theta)$ is a $N(\mu, \sigma^2)$ density, he uses $R(x^{(n)}) = (\bar{x} + k_1 s, \bar{x} + k_2 s)$, where $\bar{x}$ and s have their usual meanings and k_1 and k_2 are constants, and so achieves a disengagement from $\theta = (\mu, \sigma)$. While we would not seek to deny the reasonableness of this choice—based presumably on the fact that $(\bar{x}, s)$ is minimally sufficient for (μ, σ) and on an attempt to reflect, in the linear expressions $\bar{x} + k_1 s$ and $\bar{x} + k_2 s$, the linearity of percentiles in μ and σ—no case seems to be made out by the frequentist that there is just one such pivotal function of the minimal statistic or, alternatively, that the one chosen

is in some sense the best of such pivotal functions. Moreover, in cases where no minimal sufficient statistic exists, it is difficult to see on what principle, other than that of expediency, the frequentist could choose his pivotal statistic.

We shall see later, in Section 5, that the frequentist's difficulties in the three aspects of formulation, interpretation and mathematical development grow when he is faced with a situation, where there are certainly many possible R which satisfy his requirements and he has to incorporate some additional optimality criterion in his choice of R.

We can express the frequentist approach in terms which allow an easier comparison with the Bayesian formulations. For we can assign a utility $U(R, \theta)$ to the choice of a tolerance region R when the parameter value is θ in the following way:

$$U(R, \theta) = \begin{cases} 1 & \text{if } P(R|\theta) \geqslant c, \\ 0 & \text{if } P(R|\theta) < c. \end{cases} \tag{5}$$

This is simply an expression of the negative of a loss function in frequentist decision theory. The frequentist then wishes to choose his tolerance-region statistic $R(.)$ in such a way that the expectation, with respect to the $P^{(n)}(.|\theta)$ measure, of this utility is q. This is so since $U\{R(.), \theta\}$ is merely the indicator function of the set $G(\theta, c, R)$ and so

$$\int_{\mathcal{X}^n} dx^{(n)}\, U\{R(x^{(n)}), \theta\} p^{(n)}(x^{(n)}|\theta) = P^{(n)}\{G(\theta, c, R)\}.$$

3. The Bayesian Formulations

A number of statisticians—in particular, Lindley (1961), Raiffa and Schlaifer (1961) and Savage (1962)—have vigorously advocated a rethinking of statistical theory and practice in Bayesian terms. I have no wish here to enter into the general controversy over the relative merits of frequentist and Bayesian statistical analyses. My intention is rather to point out that the Bayesian approach must be regarded as a serious alternative to, or at least a useful complement of, the frequentist one in tolerance-region problems.

The Bayesian bases his action on two components—first, his posterior measure $\Pi(.|x^{(n)})$ and secondly, his *utility function* U, a real-valued function defined on $\mathcal{B} \times \Theta$ with typical value $U(R, \theta)$, the utility of choosing tolerance region R when the parameter value is θ. His action is to choose R so as to maximize his expected utility, the expectation being taken with respect to the $\Pi(.|x^{(n)})$ measure; that is, he maximizes

$$W(R, x^{(n)}) = \int_{\Theta} d\theta\, U(R, \theta)\, \pi(\theta|x^{(n)}) \tag{6}$$

with respect to R.

Stated in these general terms, a tolerance-region problem can be described as a decision problem in which *the decision space is the whole or part of the event space* $\mathcal{B}$, the decision function or tolerance-region statistic being a measurable function with domain $\mathcal{X}^n$ and range $\mathcal{B}$. Underlying the selection of a tolerance region $R \in \mathcal{B}$ there is usually the suggestion that R will contain a high proportion of the outcomes of future replicates of $\mathcal{E}$ or most of the "important" outcomes of future replicates. The choice of region should thus depend on our assessment of the advantage or disadvantage associated with R in relation to each possible outcome x. This dependence can

in fact be given full expression in our choice of utility function U. For suppose that, corresponding to each x, it is possible to assign quite a specific advantage or *value* $V(R, x)$ attaching to R. Then we would wish to choose R so as to maximize the expectation of $V(R, .)$ with respect to the marginal distribution of x, assessed in the knowledge that we have observed the outcome $x^{(n)}$ in the n-replicate experiment. This marginal distribution has density

$$g(x|x^{(n)}) = \int_{\Theta} d\theta\, p(x|\theta)\pi(\theta|x^{(n)}) \quad (x \in \mathcal{X}). \tag{7}$$

Thus we require to maximize

$$\int_{\mathcal{X}} dx\, V(R, x) g(x|x^{(n)}) = \int_{\Theta} d\theta\, \pi(\theta|x^{(n)}) \int_{\mathcal{X}} dx\, V(R, x) p(x|\theta) \tag{8}$$

with respect to R, and so we arrive at the maximization of (6), taking

$$U(R, \theta) = \int_{\mathcal{X}} dx\, V(R, x) p(x|\theta) \tag{9}$$

as the utility derived from the value function V. The utility specification therefore leads to as wide a class of problems as the value specification.

The problem treated by Raiffa and Schlaifer (1961) in their Example 6.1.2 has a V specification. If the daily demand of a perishable commodity falls outside the chosen tolerance interval $(-\infty, r)$ the supplier loses the potential sale of $x-r$ units, say at a loss $a(x-r)$; if demand falls inside the interval then $r-x$ units are lost through deterioration at a loss $b(r-x)$, say. Thus they take

$$V(R, x) = \begin{cases} -a(x-r) & (x \geqslant r), \\ -b(r-x) & (x < r). \end{cases} \tag{10}$$

While a detailed specification of V and U may be possible in many problems there are cases where $V(R, x)$ depends only on whether x falls inside or outside R and not on the actual value of x. For example, the main elements of concern to an engineer, who chooses the design of an electrical supply system so that only those demands x which lie in R are met, may be the capital cost of R and the possibility of the system failing through demand exceeding supply, with the consequent costly repair to the system and his prestige as designer; see, for instance, the example of Section 5. Another problem of demand-and-supply type is the choice of the height of the lower deck of a double-decker bus. The supply value is the height r provided and a typical demand is the height x of a standing passenger. All passengers with $x < r$ are accommodated comfortably and it is difficult to apportion different degrees of discomfort for passengers with $x \geqslant r$. Again, if a decision is to be taken, not by a single Bayesian but by a group of Bayesians each with his own utility and prior density functions, it may be necessary for the group, as a compromise measure, to adopt a V specification of the type just discussed, namely

$$V(R, x) = \begin{cases} B(R) & (x \notin R), \\ C(R) & (x \in R). \end{cases} \tag{11}$$

This leads, by (9), to the equivalent U specification

$$U(R, \theta) = A(R)P(R|\theta) + B(R), \tag{12}$$

where $A(R) = C(R) - B(R)$.

The important characteristic of the utility function defined by (12) is that it depends on θ only through the cover $P(R|\theta)$ provided by R at θ. Any utility function displaying this characteristic will be said to lead to a *restricted* tolerance region problem. It should be observed that while (12) gives the most general form of restricted U arising from a V specification there are other forms of restricted U, for example, that given by (5). The restricted form of the problem is the one traditionally associated with the branch of statistics known as tolerance-region theory.

The mathematical problem associated with either the general or restricted formulation above is, relative to the frequentist approach, the simple task of maximizing a function. The technique of maximization will depend largely on the form of U or V and, since the choice of this in any particular application requires specialized knowledge, we shall not pursue this essentially computational aspect further. Later, in Section 5, we shall study some interesting aspects of the restricted problem and its frequentist counterpart. In the remainder of this Section we return to a comparison with the frequentist treatment of Section 2.

The restricted Bayesian approach is not practicable with the utility given by (5) since clearly the maximum utility of 1 occurs when $R = \mathscr{X}$. If we therefore compromise by deciding to accept an expected utility of q near 1, so that we choose R to satisfy

$$W(R, x^{(n)}) = q, \tag{13}$$

then we obtain the Bayesian formulation closest to the frequentist. A useful alternative view of this formulation is provided in the following way. For given cover c and region R, denote by $H(c, R)$ the subset of Θ consisting of parameter values at which R gives cover at least c; that is,

$$H(c, R) = \{\theta : P(R|\theta) \geqslant c\}. \tag{14}$$

Then, since $U(R, .)$ is the indicator function of the set $H(c, R)$, relation (13) states that R is to be chosen so that

$$Q(R) = \Pi\{H(c, R)|x^{(n)}\} = q. \tag{15}$$

Our Bayesian rationale may then be expressed as follows. If we knew the true parameter value θ_0 we would have no difficulty in deciding on an appropriate R with cover c at θ_0. Since we do not know θ_0 we must make our choice taking account of our opinion or knowledge at the time of choice. After we have experimented and observed $x^{(n)}$ this is provided by the measure $\Pi(.|x^{(n)})$ associated with Θ. By choosing R so that $H(c, R)$ has $\Pi(.|x^{(n)})$ measure q we are thus, as Bayesians, saying that we are strongly of the opinion that the unknown parameter value is such that R provides at it cover at least c; the strength of the opinion is measured by the quality $q = Q(R)$. This seems a reasonable resolution of the problem and is, we believe, the kind of reasoning which takes place in the minds of many users of tolerance regions. We shall see in Section 4 that the formulation based on (15), with a choice of prior density which in a sense describes a display of ignorance about the parameter, leads to exactly the intervals quoted by the frequentist school.

The customary advantage of the Bayesian approach can now be seen. It substitutes for the choice of a whole *function* R defined on $\mathscr{X}^n$ in the face of entanglement with the parameter θ, the choice of a *region* R for the particular $x^{(n)}$ observed, free from any complication with unknown parameter values. The price that has to be paid for this great increase in mathematical tractability is the admission into the argument of prior densities. Now such an admission may be anathema to the frequentist, but in tolerance-region problems it is how I believe many decision-makers act. An engineer, faced with the design of a supply system as in Section 2, will, if he cannot experiment, behave as a rather pessimistic Bayesian and arrive at his design value by some weighing up of the likely values of θ. Indeed, this is the kind of information which is sometimes laid down for him in his guide books. In recent conversation with some engineers the view was put that it was widely recognized that the prior weights deducible from these books were often ridiculously pessimistic and based on the flimsiest of evidence, but that engineers acted by the book because this was their safeguard in law. Perhaps one method of breaking away from this situation is for the progressive engineer to use this quoted information as a prior density function, to carry out any possible experiments of observing demand and to base his design on a Bayesian tolerance region. In the unlikely event of his being called to account for his choice of design it would be up to the Bayesian to appear as expert witness for the defence. It is indeed difficult to see how else it would be sensible to combine the guide-book information and the experimental evidence other than through a Bayesian analysis.

4. Bayesian Tolerance Intervals

4.1. *A General Result*

We now investigate the case where $\mathscr{X}$ is all or part of the real line. Our main purpose is to derive, for some standard distributions with associated reasonably rich families of prior densities, Bayesian tolerance intervals as defined by (15); and further, to show their relationship to the corresponding frequentist intervals. An excellent account of criteria for the choice of suitable "conjugate" prior densities, together with the main properties of these families in relation to the experimental density, is given by Raiffa and Schlaifer (1961, Chapter 3) and there is no need to reproduce their advocacy here. We catalogue our results by way of the experimental density.

Before we consider these special densities, however, we can derive a simple result for determining an upper tolerance limit $r = r(x^{(n)})$ for the case where θ is a real parameter. Let us define by $d_c(\theta)$ the c probability point of the $p(x|\theta)$ distribution, so that

$$\int_{-\infty}^{d_c(\theta)} dx\, p(x|\theta) = c \tag{16}$$

and let $\delta_c(x^{(n)})$ be the c probability point of the $\pi(\theta|x^{(n)})$ distribution. We then have that

$$\Pi\{\theta : \delta_{1-q}(x^{(n)}) \leqslant \theta | x^{(n)}\} = q,$$

so that, if $d_c(\theta)$ decreases as θ increases,

$$\Pi[\theta : d_c\{\delta_{1-q}(x^{(n)})\} \geqslant d_c(\theta) | x^{(n)}] = q.$$

Thus

$$r(x^{(n)}) = d_c\{\delta_{1-q}(x^{(n)})\} \tag{17}$$

gives an upper tolerance limit of cover c and quality q. The corresponding limit when $d_c(\theta)$ increases as θ increases is

$$r(x^{(n)}) = d_c\{\delta_q(x^{(n)})\}. \tag{18}$$

4.2. *Upper Tolerance Limits for the Gamma Distribution*

Here

$$p(x|\theta) = \theta^k x^{k-1} e^{-\theta x}/\Gamma(k) \quad (x>0), \tag{19}$$

so that

$$d_c(\theta) = 2\chi^2(2k; c)/\theta, \tag{20}$$

where $\chi^2(2k; c)$ is the c probability point of a $\chi^2(2k)$ distribution. For this case $d_c(\theta)$ decreases as θ increases and so the limit (17) applies. We use here a gamma prior density

$$\pi(\theta) = b^a \theta^{a-1} e^{-b\theta}/\Gamma(a) \quad (\theta>0), \tag{21}$$

which results in a gamma posterior density

$$\pi(\theta|x^{(n)}) = (b+z)^{a+nk} \theta^{a+nk-1} e^{-(b+z)\theta}/\Gamma(a+nk) \quad (\theta>0),$$

where $z = x_1+\ldots+x_n$. Hence

$$\delta_{1-q}(x^{(n)}) = 2\chi^2(2a+2nk; 1-q)/(b+z)$$

and so an upper tolerance limit is

$$r(x^{(n)}) = d_c\{\delta_{1-q}(x^{(n)})\} = (b+z)\chi^2(2k; c)/\chi^2(2a+2nk; 1-q). \tag{22}$$

Lower tolerance limits may be similarly obtained. The corresponding frequentist limit, based on the sufficiency of z for θ, can be shown to be

$$z\chi^2(2k, c)/\chi^2(2nk; 1-q). \tag{23}$$

Note that this corresponds to the case $a = b = 0$ of the Bayesian result. This has the appearance of being very reasonable, for $E(\theta) = a/b$ and $\text{var}(\theta) = a/b^2$ and, letting $a\to 0, b\to 0$ so that $a/b\to\mu$, we have a limiting density with finite mean and infinite variance, which signifies considerable ignorance about θ. For the dangers of such interpretations of vagueness about parameters, however, see Raiffa and Schlaifer (1961, Section 3.3.4).

4.3. *Upper Tolerance Limits for the Normal Distribution*

For this case $\theta = (\mu, \sigma)$, $\Theta = \{\theta: \sigma>0\}$ and $p(x|\theta)$ is the $N(\mu, \sigma^2)$ density. We define $\bar{x}$ and s in the usual way by $n\bar{x} = \Sigma x_i$ and $(n-1)s^2 = \Sigma(x_i-\bar{x})^2$. The conjugate prior density is of normal-gamma type with

$$\pi(\mu, \sigma) \propto (1/\sigma)\exp\{-\tfrac{1}{2}b(\mu-a)^2/\sigma^2\}(1/\sigma)^w \exp(-\tfrac{1}{2}wv/\sigma^2); \tag{24}$$

we then say that $\pi(\mu, \sigma)$ is a $N\Gamma(a, b, v, w)$ density. It follows very simply that $\pi(\mu, \sigma|x^{(n)})$ is a $N\Gamma(A, B, V, W)$ density, where

$$B = b+n, \quad A = (ba+n\bar{x})/B, \tag{25}$$

$$W = \begin{cases} w+n & (b>0), \\ w+n-1 & (b=0), \end{cases} \tag{26}$$

$$V = \{wv+ba^2+(n-1)s^2+n\bar{x}^2-BA^2\}/W. \tag{27}$$

Since $d_c(\theta) = \mu + \nu_c \sigma$, where ν_c is the c probability point of the $N(0, 1)$ distribution, we require by (15) to choose r so that

$$\Pi\{(\mu, \sigma) : \mu + \nu_c \sigma \leqslant r \,|\, x^{(n)}\} = q. \tag{28}$$

If we introduce new parameters (ξ, η) by

$$\xi = B^{\frac{1}{2}}(\mu - A)/\sigma, \quad \eta = V/\sigma, \tag{29}$$

then the inequality in (22) becomes

$$(\xi + \nu_c B^{\frac{1}{2}})/\eta < B^{\frac{1}{2}}(r - A)/V. \tag{30}$$

Now the posterior distribution of (ξ, η) is such that ξ and η are independent with ξ distributed as $N(0, 1)$ and η as $\{\chi^2(W)/W\}^{\frac{1}{2}}$, and so $(\xi + \nu_c B^{\frac{1}{2}})/\eta$ has a non-central t-distribution with W degrees of freedom and non-centrality parameter $\nu_c B^{\frac{1}{2}}$. If $t(W, \nu_c B^{\frac{1}{2}}; q)$ denotes the q probability point of such a non-central t-distribution then it follows from (30) that

$$B^{\frac{1}{2}}(r - A)/V = t(W, \nu_c B^{\frac{1}{2}}; q)$$

and so

$$r = A + VB^{-\frac{1}{2}} t(W, \nu_c B^{\frac{1}{2}}; q). \tag{31}$$

When $b = w = 0$ we have by (25), (26) and (27) that

$$r = \bar{x} + sn^{-\frac{1}{2}} t(n - 1, \nu_c n^{\frac{1}{2}}; q), \tag{32}$$

which is the familiar frequentist limit. The improper prior density associated with $b = w = 0$ is of Jeffreys type and is usually interpreted as a display of ignorance about the parameters μ and σ; see Jeffreys (1961, Section 3.4.1).

4.4. *Upper Tolerance Limits for Poisson and Binomial Distributions*

For discrete distributions over non-negative integers we define $d_c(\theta)$ as the smallest integer satisfying the inequality

$$\sum_{x=0}^{d_c(\theta)} p(x \,|\, \theta) \geqslant c, \tag{33}$$

and the Bayesian upper tolerance limit $r(x^{(n)})$ is defined by

$$\Pi\{\theta : d_c(\theta) \leqslant r \,|\, x^{(n)}\} = q. \tag{34}$$

For the Poisson experimental density

$$p(x \,|\, \theta) = e^{-\theta} \theta^x / x! \quad (x = 0, 1, 2, \ldots) \tag{35}$$

with gamma prior density (21) we obtain a gamma posterior density with

$$\delta_q(x^{(n)}) = 2\chi^2(2a + 2z; q)/(b + n). \tag{36}$$

Since $d_c(\theta)$, which is easily obtainable from Poisson tables, increases with θ, the limit $r(x^{(n)})$ is readily obtained by (18). Again, as in Section 4.2, the case $a = b = 0$ produces a frequentist tolerance limit.

A similar type of analysis can be applied to the binomial experimental density with a beta prior density. For difficulties in the interpretation of the case which leads to frequentist results see Raiffa and Schlaifer (1961, pp. 63–65).

4.5. *Finite Tolerance Intervals for the Normal Distribution*

Finally in this Section we consider the problem of obtaining a Bayesian tolerance interval (r_1, r_2) for the normal experimental density with prior density (24). By (15) we have to choose a pair (r_1, r_2) to satisfy the probabilistic relation

$$\Pi[(\mu, \sigma) : \Phi\{(r_2-\mu)/\sigma\} - \Phi\{(r_1-\mu)/\sigma\} \geqslant c \,|\, x^{(n)}] = q, \tag{37}$$

where Φ denotes the distribution function of a $N(0, 1)$ distribution, and the transformation (29) provides the equivalent inequality

$$\Phi\{B^{-\frac{1}{2}}\xi + (r_2-A)V^{-1}\eta\} - \Phi\{B^{-\frac{1}{2}}\xi + (r_1-A)V^{-1}\eta\} \geqslant c.$$

If $k_1 = (r_1-A)V^{-1}, k_2 = (r_2-A)V^{-1}$ satisfy the new probabilistic statement then

$$r_1 = A + k_1 V, \quad r_2 = A + k_2 V. \tag{38}$$

It can be shown that the length of the interval $(k_2-k_1)V$ is minimized when $k_2 = -k_1 = k$ say, so that $(A-kV, A+kV)$ would be a satisfactory tolerance interval. Because of the distributional properties of (ξ, η) the value of k is exactly that given in frequentist tables; see, for example, Owen (1962, Section 5.4), where $B^{-\frac{1}{2}}\xi$ corresponds to the estimator of μ based on B observations and η the independent estimator of σ with W degrees of freedom. Again the case $a = b = 0$ yields the usual frequentist interval $(\bar{x}-ks, \bar{x}+ks)$ based on the outcome $x^{(n)}$ of n replicates.

5. Optimum Tolerance Regions

On the assumption that the frequentist has chosen his pivotal statistic or the Bayesian his prior density, the tolerance-interval problem, based on (2) or (15) and developed in Section 4, usually leads to a unique sensible solution. This is not so, however, when $\mathcal{X}$ is higher-dimensional. There may then be many statistics R satisfying (2) and many regions satisfying (15) so that formulations based on the frequentist utility (5) are incomplete. A problem of this type is the so-called "diversity problem" of engineers, illustrated by the following simple example.

Example. An electrical supply system of the type illustrated in Fig. 1 is to be designed. "Operations" of the system are independent and the loads "demanded" at

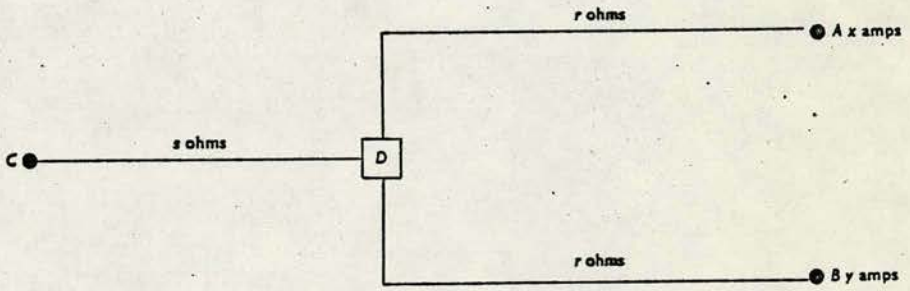

FIG. 1. Electrical supply system.

terminals A and B—of x and y amps, say—are independently and identically distributed, the variation from operation to operation being described by a measure $P(.\,|\,\theta)$ associated with the (x, y)-plane. The system succeeds at an operation if the drops in voltage between C and A and between C and B are both less than v volts. A design consists of choosing the sizes of the cables AD, BD and CD, the sizes being

most conveniently expressed in terms of the associated resistances, say r, r and s ohms. Then clearly the cover at θ provided by the *design* (r,s) is given by $P(R|\theta)$, where

$$R = \{(x,y) : rx+s(x+y)<v, ry+s(x+y)<v\}.$$

If there is available information on the potential demands on the system—possibly the observed demands on an already existing system which has been so overdesigned that failure is virtually impossible—then we might attempt to regard the choice of design (r,s) as the choice of a tolerance region R to satisfy (15), or, in the frequentist case, a tolerance-region statistic R to satisfy (2). But here there are clearly many statistics and many regions satisfying (2) and (15), respectively, and it is obvious that, to make the problem sensible practically, we must take account of the cost $K(R)$ associated with the construction and/or operation of the design (r,s). Regarded as a restricted problem, the utility must take account of both $P(R|\theta)$ and $K(R)$.

Even in the one-dimensional case it is difficult to feel satisfied with the frequentist formulation characterized by (2) or its Bayesian counterpart, for the utility $U(R, \theta)$ in a restricted problem will usually take account of two competitive factors. The more extensive the set R the greater is the cover provided by it at θ and so the greater will be the utility. On the other hand, there must usually be disadvantages attached to increasing R, such as increased cost or length of interval; if not, why, in any practical problem, do we not quote $\mathscr{X}$ as a region which certainly provides cover 1? In the remainder of this Section we term this disadvantage factor a cost $K(R)$ and find it convenient to speak in terms of the operations of systems such as the supply system of the above example.

If a cost $K(R)$—such as running cost, possibly with some allowance for capital cost or depreciation—can be assigned to each operation of the system, then it may be appropriate to consider the following restricted V specification.

$$V(R,x) = \begin{cases} \lambda_1 - K(R) & (x \notin R), \\ \lambda_2 - K(R) & (x \in R). \end{cases} \tag{39}$$

Here λ_1 is the (negative) monetary value associated with failure of an operation and λ_2 the monetary value associated with success of an operation. The corresponding U specification is, by (12),

$$U(R,\theta) = \lambda P(R|\theta) - K(R),$$

where $\lambda = \lambda_2 - \lambda_1$. The Bayesian formulation in terms of the maximization of $W(R, x^{(n)})$, as given by (6), is then straightforward. For instance, simple calculations show that, for the experimental and prior densities of Section 4.2, the region $R = (0, r)$ must be chosen so as to maximize

$$\lambda I_{r/(b+s+r)}(k, a+nk) - K(R), \tag{40}$$

where I denotes the incomplete beta-function (Pearson, 1934). This presents no computational difficulties. The frequentist, on the other hand, is faced with $-U(R,\theta)$ as his loss function in the general decision problem and hence with the formidable task of finding a *function* R which maximizes

$$\int_{\mathscr{X}^n} dx^{(n)} p(x^{(n)}|\theta)\,[\lambda P\{R(x^{(n)})|\theta\} - K\{R(x^{(n)})\}] \tag{41}$$

for every θ. It is quite possible that no such function exists and he would be forced to adopt one of the other escape routes from the entanglement with θ indicated in Section 2. It seems to be a fair generalization that the more sophisticated the specification of the utility or loss function the more the Bayesian enjoys the situation, whereas the more the frequentist encounters difficulties.

When the cost associated with the system is essentially the capital cost $K(R)$ of constructing the design or region R and there is pressure to provide cover c, then a more suitable form of utility function may be given by

$$U(R, \theta) = \begin{cases} \lambda_1 - K(R) & \text{if} \quad P(R|\theta) < c, \\ \lambda_2 - K(R) & \text{if} \quad P(R|\theta) \geq c. \end{cases} \tag{42}$$

We then have, by (6), that

$$W(R, x^{(n)}) = \lambda \Pi\{H(c, R)|x^{(n)}\} - K(R) = \lambda Q(R) - K(R), \tag{43}$$

where $\lambda = \lambda_2 - \lambda_1$ and an additive constant has been omitted. The difficulty with such a formulation is that it may be practically impossible to assess λ. In these circumstances, if the pressure to provide cover c is strong, it is interesting to study the following resolution. From the set of regions satisfying the guarantee of cover given by (15) choose one which minimizes the cost.

The frequentist, trying to formulate his corresponding problem, has first the formidable task of discovering the set $\mathscr{R}$ of all statistics satisfying (2). Even if he can find $\mathscr{R}$, how does he then proceed to choose the statistic which gives minimum cost? There is unlikely to be a statistic $R^* \in \mathscr{R}$ which provides uniformly minimum-cost regions in the sense that

$$K\{R^*(x^{(n)})\} = \min_{R \in \mathscr{R}} K\{R(x^{(n)})\}, \tag{44}$$

for all $x^{(n)} \in \mathscr{X}^n$. He may then decide to take as a suitable tolerance-region statistic that R which satisfies (44) for the particular $x^{(n)}$ he observes, but this suffers from the defect, unpopular with frequentists, that the statistic used depends on the particular $x^{(n)}$ which turns up rather than on some properties of measures over the outcome space. Clearly there are great difficulties involved even in the formulation of the frequentist problem.

The Bayesian, freed from the necessity of determining whole functions in the presence of unknown θ, is able to formulate the problem much more simply. We can, in fact, pose two different problems and exploit a useful duality property of them.

Fixed-quality q problem. We wish the selected region R to satisfy (15) so that the quality $Q(R)$ of the region is fixed at q. If $\mathscr{S}(q)$ denotes the set of all such regions, that is, if

$$\mathscr{S}(q) = \{R : Q(R) = q\}$$

we then wish to minimize $K(R)$ over $\mathscr{S}(q)$.

Fixed-cost k problem. We suppose that the cost of the selected region is to be fixed at k so that we are interested only in regions in the set

$$\mathscr{T}(k) = \{R : K(R) = k\}.$$

We then wish to choose a region $R^*(k)$ which maximizes $Q(R)$ over this set; that is,

$$Q\{R^*(k)\} = \max_{R \in \mathcal{T}(k)} Q(R) = L(k), \tag{45}$$

say. In what follows we make the assumption, reasonable in many applications, that the maximum quality $L(k)$ is a continuous and strictly increasing function of k.

Solutions $R^*(k)$ of fixed-cost problems lead easily to the solution of the fixed-quality q problem. If, in fact, we determine k^* such that $L(k^*) = q$ then $R^*(k^*)$ is an optimum tolerance region of fixed quality q; the minimum cost is k^*. For, suppose that R_1 with cost $k_1 < k^*$ and not $R^*(k^*)$ is optimum. Then we have

$$q = Q(R_1) \leqslant \max_{R \in \mathcal{T}(k_1)} Q(R) = L(k_1) < L(k^*) = q,$$

a contradiction, so that $R^*(k^*)$ is in fact an optimum fixed-quality q region.

Solutions of fixed-quality problems could, by similar reasoning, provide a solution of the fixed-cost problem. Since the set $\mathcal{T}(k)$ is usually easily defined and in no way depends on θ the fixed-cost problem is likely to be a straightforward maximization problem. From its solutions for different k the function $L(k)$ can be explored to obtain k^* and hence $R^*(k^*)$.

It is not our purpose here to explore further the computational aspects of this problem, which, for complicated systems, are considerable. Our objective has been the limited one of showing that the Bayesian approach, in contrast to the frequentist one, to the search for optimum tolerance regions leads to well-formulated problems.

6. Some General Remarks

It is my hope that this paper and that by Mr Thatcher which follows are provocative enough to lead to a discussion which will sweep tolerance-region theory out of the doldrums in which it seems to be at present becalmed. While I have no doubt that frequentists and Bayesians will try to create such a storm as to drive the craft into port, their own or their opponents'—I am not sure which—I hope that the direction of motion may be influenced by the views of users of tolerance regions. A very useful contribution to the discussion could be the opinion of users as to which type of problem—the general or the restricted type—they have met, what utility or loss functions they regard as useful and how applicable they, as users, judge frequentist and Bayesian analyses.

Acknowledgements

I am grateful to members of the Hospital Engineering Research Unit of the University of Glasgow, who, by their patient attention to my explanations of frequentist tolerance-region theory and by their reluctance to accept its concepts and interpretation, first drew my attention to the need for an alternative approach. I should also like to thank both referees for their helpful, constructive criticism.

References

Bowker, A. H. (1947), "Tolerance limits for normal distributions", in *Selected Techniques of Statistical Analysis*, edited by Eisenhart, C., Hastay, M. W. and Wallis, W. A., Chapter 2. New York: McGraw-Hill

Fraser, D. A. S. (1953), "Non-parametric tolerance regions", *Ann. math. Statist.*, **24**, 44–55.

—— (1957), *Nonparametric Methods in Statistics*. New York: Wiley.

Jeffreys, H. (1961), *Theory of Probability*. Oxford University Press.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*. New York: Wiley.

LINDLEY, D. V. (1961), "The use of prior probability distributions in statistical inference and decision", *Proc. 4th Berkeley Symp.*, **1**, 453–469.

OWEN, D. B. (1962), *Handbook of Statistical Tables*. London: Pergamon.

PEARSON, K. (1934), *Tables of the Incomplete Beta-Function*. Cambridge University Press.

RAIFFA, H. and SCHLAIFER, R. (1961), *Applied Statistical Decision Theory*. Boston: Graduate School of Business Administration, Harvard University.

SAVAGE, L. J. (1954), *The Foundations of Statistics*. New York: Wiley.

—— (1962), *The Foundations of Statistical Inference*. London: Methuen.

WALD, A. and WOLFOWITZ, J. (1946), "Tolerance limits for a normal distribution", *Ann. math. Statist.*, **17**, 208–215.

WALLIS, W. A. (1951), "Tolerance intervals for linear regressions", *Proc. 2nd Berkeley Symp.*, **1**, 43–51.

WEISSBERG, A. and BEATTY, G. H. (1960), "Tables of tolerance-limit factors for normal distributions", *Technometrics*, **2**, 483–500.

WELCH, B. L. and PEERS, H. W. (1963), "On formulae for confidence points based on integrals of weighted likelihoods", *J. R. statist. Soc.* B, **25**, 318–329.

The preceding paper was read at a Research Methods Meeting of the Royal Statistical Society on 5 February 1964. At the same meeting a paper on a related subject was presented by A.R. Thatcher. The discussion and replies which customarily follow such read papers therefore refer to both these papers.

DISCUSSION ON THE PAPERS BY MR AITCHISON AND MR THATCHER

Mr C. B. WINSTEN: We have just heard two very interesting papers. I feel some diffidence in opening the discussion on them for at least two good reasons. One is that Mr Aitchison throws out an invitation to the users of the statistical techniques he has just discussed. In this particular discussion, though I cannot really claim to be a user of tolerance intervals, I shall have to present myself perhaps as a theorist for them, but I very much hope that real users will join in as he suggests. Mr Aitchison gives thanks to users who have been sceptical and forced him to reconsider his notions. What a very healthy state of affairs this is. We are all realizing that the diversity of problems to which

statistical methods are applied have many subtle logical differences, and that these differences should not only be understood by the statistician but should be reflected in his treatment of the problem and in his interpretation of the results. Unfortunately we have been rather type-cast in the past as statisticians because our modes of reasoning have been rather limited and have been inflicted on the users rather than arising from their demands. So we must give special thanks to sceptical users, such as the Glasgow Hospital engineers that Mr Aitchison mentions, and also to responsive statisticians, such as Mr Aitchison, who respond to the points that these users put and try to ensure that the statisticians' logic fits the situation.

My other reason for being short is that I am sure, with a high probability, though one that I would not like to give a number to, that yourself, Mr Chairman, and other Bayesians are just aching to join in this discussion.

Mr Aitchison's paper, to which I want to address my remarks in the most part, is a very lucid one and, like very many lucid ones, it arouses some questions, and indeed misgivings, about the problems that he is discussing, and their formulation. Both papers, I think more Mr Aitchison's, acknowledge these tolerance interval problems basically as *decision* problems. If one considers, for example, production of a particular item, intervals are simply an aid on the way to a decision, perhaps help on the way to a decision for a large number of different people with different uses. If one gives the prediction of how productivity is going to be next year, or something of that sort, then one is usually giving the prediction to help people who are guiding the economy, i.e. help people who are going to have to make decisions. So I accept the fact that these problems basically are decision problems, ones with particular sorts of risk functions, as Mr Aitchison points out in his paper, that have basically two values, though there can be differences.

What sort of special cases can there be for the use of this sort of decision problem? Let me try to give two sorts. The first one might be something like this. A manufacturer is to mass produce a device. He takes a sample of conditions under which the device will have to function, and then fixes his tolerances as the result of this sample, and goes into full production. He cannot then change his production after he has started. Some of his devices meet extra conditions outside the tolerance interval and they fail. The proportion that fail may be a plausible measure in this particular case, and this is the point I want to emphasize, of the cost of setting up the process in the way that he did. And the proportion that do not fail correspondingly is a measure of the success. The light-hearted example that Mr Aitchison gives of people in a bus is really a case just like this. But there are really quite serious examples which are similar, the sort of ergonomic ones where in fact people do not know all the statistics of the users. In the case of the bus, of course, the bus designer would only have to look up statistics for the whole population—we hope he would not just take a sample. But there are many cases where this sort of thing would happen. Now in this sort of example the idea of a cover which Mr Aitchison has introduced occurs not as a probability but basically as a utility, as a payoff function. The payoff may in some cases be uncertain because the initial sample on which the tolerance interval is based is finite. In such a case one has great misgivings about the whole notion of the quality that has been introduced. Why should one fix so severely on a single value c for the payoff function? Why concentrate on getting a probability, in the Bayesian formulation, of getting the payoff above c? What one should be interested in is looking at the whole payoff or cover function, as it depends on θ and, of course, on the fixed sample that one took at the beginning. The fixed sample has been taken, so basically one will simply have a cover function as a function of θ. And then what one should do is to consider this payoff function in the light of the likelihood function (or the posterior probability function, if one were a very convinced Bayesian), obtained from the initial sample that one had got. Now it may be objected that this sort of notion is not a true tolerance interval problem. But I think it does have all the characteristics of the sort of problem that Mr Aitchison is thinking of. Another type of problem that perhaps comes

more closely to the spirit in which these things have been introduced is one of a sort, again the sort that Mr Aitchison has defined, where one might have a reservoir and have to define the size on the basis of a fixed sample of rainfalls. Or one might have a storage system and basically only have a limited sample in which this storage system is used. Again this reservoir will only have a life of so many years imputed to it, so one would be interested in the finite sample that the reservoir will face after it has been built. In this particular case cover is not a *payoff* function, it is a *probability*, but I think the same sort of point comes up. Basically one would be interested in plotting the cover as a function of θ after one has taken one's first sample. I would want the cover to be high where the likelihood or the posterior probability, crossing fingers as before, was high, and only low where all the evidence pointed to an infinitesimal chance of the parameter being the true one. Incidentally, Mr Aitchison raises the notion of a statistician having to give evidence in a law case, and if I were giving evidence, or advising in this situation, I would be very sceptical of the notion of producing a Bayesian prior probability distribution. What I would expect to do would be for the statistician to produce his cover function, and if the cover function was low in a particular range of values of θ, say, I would expect him not merely to say that he had a high prior probability of values excluding this particular portion where the cover was low, but to show evidence that he had really thoroughly investigated these values of θ to make sure that that sort of value of the parameter had not actually occurred. Quite often, in the real world, as against the very formalized world of these tolerance intervals, there is lots of ancillary evidence or extra experiments that one can go to to see what sort of value of θ there might be. And it may be perfectly possible to follow up these particular values of θ and see that they could not really happen at all. That is not, of course, to say that the more mechanical sort of Bayesian reasoning is not useful for deriving a procedure, of course it is. The very best way, as we know from many examples, for getting a particular procedure is to try out a prior probability distribution, get a procedure. What I am saying is that having got this procedure one should plot the whole cover function in a problem of this sort.

I was worried, too, in this light, at a remark that Mr Thatcher makes on page 188 of his paper. There he states that he gives an idea of a mixed strategy. He says, just above 7.2, "Similarly, under the Bayesian approach, there may be occasions when the prior knowledge is somewhat imprecise, so that one hesitates between one prior distribution and another." Again it will be possible to mix the Bayesian solutions by a random experiment. I think in that particular case what one would be more likely to do is to put the plausible prior distributions, ones that you might think apply in extreme cases, and so on, produce solutions for them, and then plot the cover functions, to see whether in fact they had this sort of dip in unfortunate places, and so on. I do not think in the least that one would simply carry on straightaway and get a unique solution in this sort of thing without checking on the cover function or the risk function first.

The essential thing is basically to keep a distinction between those parts of the problem that can be found by calculation, and this would be the cover function, and the subjective elements where they are brought in, to demonstrate entirely, especially if one has to argue with a variety of different users, or even legal minds, just how these things are to be kept distinct.

I found both these papers very interesting ones, as I said, and I have much pleasure in moving the vote of thanks.

Mr D. F. Kerridge: I rather imagine that I was asked to second the vote of thanks because it was feared that the speakers might not be controversial enough, and I will do my best to raise still further any temperatures that have not already reached boiling point. But first I must praise Mr Aitchison's paper on two points which I think are very important. First, that it looks as if it is meant to be read. So many papers look as if they are meant to be filed. Secondly, that he recognizes explicitly that what we do, or rather what we as theorists recommend that others do, must be determined by the needs of the user. You

may remember that Dr Good at one of these meetings (Good, 1960) suggested that operations research is needed to determine what users actually do with the answers we give them. We theorists have in the past paid very little attention to this. The question then arises: what do users really want? Now it may very well be that users do not want a tolerance interval. In fact, given the right situation, nobody would want it. If you have got understandably defined costs, and you have decisions to make, I do not think anybody would dream of treating it in any other way than as a decision problem. This means, as we know, that we must use Bayesian strategies, so I do not think that Mr Aitchison is really saying anything which is very controversial here, from whichever point of view you like to look at it. Nobody has yet suggested that decision theory is wrong for the kind of problem in which you are making decisions. I am therefore in complete accord with the practical recommendations of the paper. However, reading between the lines to find something controversial, I find the suggestion that what we are trying to do is to tell the engineer what he ought to believe. Now I have no great experience with engineers, but I do not think an engineer wants to be told what he ought to believe. He wants to be told what he can do that will fairly certainly be right. It is true, of course, that very often Bayesian methods do lead us to make decisions which are right, but there seems to be some suggestion that there is some other reason for using them. I claim that the only respectable reason for using Bayesian methods is because you are a frequentist. Now I know that Bayes's theorem fell into disfavour for a short time among frequentists, although von Mises himself, the most vigorous advocate of the frequency theory, consistently maintained that the use of Bayes's theorem was the only sensible way to make an inference. I think that this view is one which is likely to return. If we look at the justifications which are given of the use of Bayesian inference from the non-frequentist point of view, the usual explanation given is that these are the rules that you must adopt in order to be consistent. Now I do not know that anybody wants to be consistent. After all, we can be consistently wrong, just as easily as consistently right. I admit the appeal of consistency, but the appeal is only to our aesthetic instinct for tidiness, which is not necessarily a safe guide in practical affairs. The engineer, I should imagine, is not concerned with being consistent, he just wants to be right. This brings me back to my main thesis that the right reasons for using Bayes's theorem are frequency reasons. The difficulty is perhaps that the frequency properties of Bayesian solutions are not really well understood yet, though we are getting to understand them more. We have here in these two papers practical examples in which Bayesian and confidence solutions agree to within any reasonable practical degree. Take Mr Thatcher's paper: the difference between the Bayesian and confidence interval solutions is of the order of one observation, which is neither here nor there in any practical situation. We may like, as theorists, to have exact solutions, but it does not really bother the practical man very much. Nothing fundamental can depend on one observation. And, in fact, the randomized strategy Mr Thatcher suggests in the final part of the paper indicates even more strongly that the difference between the confidence interval and the Bayesian solution is very small, because he finds that by taking a mixture of the two we can get randomized confidence intervals which have slightly greater efficiency than the ordinary confidence solution. The difference between the two solutions in the first place was only one observation either way: by taking a half-and-half mixture of the two things you have got an even smaller difference. Obviously, if you took some prior distribution which was half-way between the two you would get inferences which in the long run would give you sensible frequency answers. There are already a few results known on the frequency properties of Bayesian inferences. I myself proved a little while ago (Kerridge, 1963) that given a finite number of hypotheses you cannot often end up with the wrong result if you apply Bayes's Postulate. And I understand that Professor Savage has a result which shows that in large samples you must have agreement between confidence intervals and the results of applying Bayes's theorem and constructing intervals from the posterior distribution. Clearly, there is a great deal of

further investigation needed, but this seems a very hopeful line of development. What we really need are rules for finding inoffensive prior distributions which will make the convergence between the two kinds of solution even more rapid, such as those given by Dr Welch and Dr Peers in their recent paper in Series B (Welch and Peers, 1963). To a first approximation it is fairly easy to see what to do. For instance, if you take the case where you have discrete hypotheses you find that the posterior probability of each hypothesis is roughly normally distributed in large samples, by a simple application of the central limit theorem. If you arrange that the prior probability distribution is such that the means of these distributions are equal then you are going to get the convergence between the two methods pretty rapid. And this seems the only really satisfactory way to choose prior distributions, that is, to choose them so that they give sensible answers. In fact, I think that this has already been applied to some extent. If you look at the ways in which Jeffreys chooses prior distribution (Jeffreys, 1961, p. 192) he does seem to bear very much in mind that they should lead to reasonable posterior distributions. Or at least he produces a rule by, say, invariance theory which gives him a prior distribution and then rejects it because he does not like the answers it gives. Far from being illogical in doing this, I think this is precisely the right thing to do. And this, I hope, is what we shall see happening in the future.

I must thank both the speakers very much. I wish they could have said something more controversial. I have much pleasure in seconding the vote of thanks.

The vote of thanks was carried unanimously.

Professor E. S. PEARSON: I am glad that Mr Thatcher has brought out again for inspection this historic problem of the two binomial samples, because each generation of statisticians responds to the outlook of its time by giving a rather different twist to the possible solutions. There is some progress, a greater understanding of the relationship between the alternatives, but to me at any rate it seems doubtful whether there can ever be a uniquely acceptable answer.

I am not going to be drawn into the heat of controversy by Mr Aitchison's challenging remarks, but should perhaps go this far in a confession of faith. For myself—and I want to emphasize that this is a purely personal attribution—I feel a certain dishonesty in inserting into the foundation of my mathematical inference structure a function, say $f(p)$, which will often have no real meaning for me. In reaching conclusions on the basis of statistical data there are many factors to be taken into account which do not seem to be expressible in terms of numbers, and prior knowledge may often for me have to be one of these. But to quote a recent remark of M. G. Kendall's, "a man's attitude towards inference . . . is determined by his emotional make-up, not by reason or by mathematics"; perhaps I might add also, by the conditioning of his life experience.

In this connection, then, may I be allowed to make a few remarks from the historical angle as to where Mr Thatcher's problem stood in 1922 when Dr Irwin, our President, and I happened to be learning our statistical theory together at University College. First, following a good tradition, I will remark in passing that his equation (16) was given in a paper of mine published in *Biometrika* in 1925, except that in my particular approach the prior distribution of p was taken as symmetrical with $b = c$ in his equation (15). I am not suggesting that you should read my paper which, by modern standards, was not perhaps a very good one, being far too discursive. I could make much better use of the data now! But it is relevant, I think, to refer to it here as one outcome of discussions which were taking place at a point in time over 40 years ago when ideas on inference were beginning to be thrown into the melting pot.

What was the position then? The early British school of applied statisticians had not seemed to feel the need to think very deeply about the basis of inference; on the whole their outlook was certainly a frequentist one, but at times they felt compelled to appeal to Bayes's theorem. Thus, my father in 1920 wrote a paper called "The fundamental

problem of practical statistics", which starts from Bayes's theorem. He thought he had found a solution to the dilemma but in this I am sure he was wrong. Starting in terms of Bayes's illustration of the balls on a billiard table, he pointed out that if the probability distribution governing the position of the later balls was the *same* as that for the first, marker ball, the posterior distribution for Mr Thatcher's a_2 assumed Laplace's form whether the prior distribution $f(p)$ was uniform or not. This assumption of a common distribution might be true for dropping balls onto a table, but in general, of course, is not likely to hold.

My own investigation must, I fancy, have been stimulated by K. P.'s reference to some remarks of Edgeworth made as long ago as 1884. Writing in the journal *Mind*, Edgeworth had asserted:

> I submit, the assumption that any probability-constant about which we know nothing in particular is as likely to have one value as another is grounded upon the rough but solid experience that such constants do as a matter of fact, as often have one value as another (p. 230).

And again:

> We take our stand upon the fact that probability-constants occurring in nature present every variety of fractional value; and that natural constants in general are found to show no preference for one number rather than another. Acting on which supposition, while in particular cases we shall err, in the long run we shall find our account (p. 231).

You will see that Edgeworth is appealing not to an equal distribution of ignorance nor to a subjective measure of probability associated with a particular set of circumstances but to a loosely defined, general statistical experience. It was to examine this claim that I collected a very large number of twin binomial samples, n_1, n_2, from many different kinds of population. I then examined how far the frequency a_2 in the second sample fell within the Bayes's posterior limits,

(1) taking $f(p)$ as constant,

(2) giving it U- or V-shaped forms corresponding roughly to the actual distribution of $(a_1+a_2)/(n_1+n_2)$ found in my particular large-scale statistical experience.

In this way it may be said that I started my statistical career as a Bayesian frequentist! I am interested to hear that this is what Mr Kerridge thinks that one should be!

But while its consequences could be explored in this simple classical problem, a Bayesian frequentist attack of this kind did not then seem to be a very profitable line of study. And at this juncture came Fisher's 1922 Royal Society paper with its condemnation of the use of inverse probability. It was hardly surprising that most of the statisticians of my generation—and of course there were not many of us—who had certainly not felt happy with Bayes, moved away (as Fisher himself admitted he had done) from the Bayesian tradition which had still had some influence on our elders.

As a result, an interesting consequence followed. With the use of posterior distributions ruled out of court, there was a strong incentive to find some method for providing interval estimates for unknown parameters in the case of small samples, a field in which Fisher's work had provided so many distributions. To meet this need, the fiducial and confidence interval theories were evolved independently but almost simultaneously in the late 1920's.

Now, of course, in the 1960's, these problems are looked at with revived interest in the light of the subjective theory of probability. We understand more and these discussions are a sign that we are alive and healthy. But there is always a question in my mind. Is any advocate of the Bayesian approach yet prepared to take responsibility for writing a new *Statistical Methods for Research Workers* for the plain working statistician?

Mr T. W. MAVER: The engineer's task in the design of physical systems is twofold: he must assess the demand which will be made on the system and he must provide a system which is physically capable of meeting this demand.

In undertaking the first of these tasks, an activity for which he is academically ill-prepared, he may encounter a loading distribution in which the maximum demand is so large as to be indeterminable, and thus be obliged, albeit with a clear conscience, to accept the inevitable inadequacy of his design, or he may encounter a loading distribution in which the maximum demand may conceivably be met, and thus be presented with a dilemma based on the understanding that total satisfaction is within his power, if not his pocket.

In either event, from the outset, an understanding of the real world interpretation of "cover" and "quality" is necessary. Although the engineer may understand the relative frequency interpretation of "P" he is unable, even after consultation with the user, to arrive at a sensible measure of cover, let alone quality, in ignorance of the consequences of design failure.

Moreover, his decision regarding the level of the design cannot sensibly be made, whether the outcome space is uni- or multi-dimensional, in conscious or unconscious repudiation of his obligation to the possibly dissatisfied users of other unsatisfactory systems which function serially or coincidentally with the system under investigation.

The opportunity afforded by a Bayesian approach—of admitting a measure of the consequences of design failure and a measure of the utility of the operation—is an attractive one. In many cases, however, the physical ramifications of failure may not be readily determinable while in other cases, although all the physical aspects are understood, their effect on the user may defy simple quantitative assessment.

Although the engineer has been in the past largely unaware of the wealth of statistical help on which he could draw to assist him in the definition of loading distributions, he has never been unmindful of the abnormal occurrences, due either to the malfunctioning of a piece of equipment or to the stupidity on the part of the user, which would bring about design failure. The development of automatic control devices, pursued with the object of obviating the dangers which might accompany such abuses, has, fortunately, if inadvertently, brought about a simplification in the specification of the value function "V". This can best be illustrated by an example.

A fusible link is incorporated in an electrical circuit supplying a number of socket outlets; any demand in excess of that for which the circuit is designed effects a break in the circuit by fusing the link and the system becomes inoperative. The fact that the resulting inconvenience, in terms of the trouble to repair the fuse and the lack of facilities during the repair period, is the same whatever the magnitude of the excess demand, indicates that this situation is of the restricted type. A more sophisticated form of control, such as might be found in an elevator, wherein an automatic reset is actuated when the excess load is removed constitutes a "V" specification in which the elements of cost due to inconvenience are replaced to a large extent by the capital cost of the control mechanism in the evaluation of λ_1.

Although most engineering problems fall into the restricted category there are some which do not. If we consider the supply of hot water to a hospital ward unit we encounter a most intractable situation. Failure of the design obviously results in some inconvenience to the staff and possible danger to the patient. The degree of this inconvenience and danger is dependent on the form in which the failure manifests itself: it may be reduced flow of hot water, it may be slight reduction in temperature over a protracted period or it may be drastic reduction in temperature over a short period. This lack of understanding of the physical possibilities together with the acutely difficult task of evaluating inconvenience and danger presents such a complex problem that the engineer may be glad to revert to the vagueness of Frequentist interpretation.

In any event, when the storm of discussion has finally abated, the engineer will be quite happy if he discovers the craft safely moored in either one of the two ports rather than at the bottom of the sea.

Dr J. A. HARTIGAN: I would like to make one or two purely technical comments. First, it seems reasonable to distinguish completely between the observation space $\mathcal{X}$

and the prediction space $\mathscr{Y}$, which are connected by having probability distributions indexed by Θ. For example, in a regression problem, $\mathscr{X}$ would represent observations for certain values of the fixed variables, and $\mathscr{Y}$ would represent a new observation on a new set of values of the fixed variables.

The tolerance region R_x in $\mathscr{Y}$ (having observed x) is of particularly simple form if Mr Aitchison's (39) is such that the cost $K(R_x)$ is a positive measure, with density $k(y)$, say. Then the Bayesian procedure with prior density $\pi(\theta)$ gives

$$R_x = \{y; \pi(y \mid x) > k(y)\}.$$

The asymptotic behaviour of regions of this form is accessible if $k(y)$ is a prior density on Y corresponding to a prior density $l(\theta)$ on Θ; mathematically, if

$$k(y) = \int_\Theta p(y \mid \theta)\, l(\theta)\, d\theta.$$

We can assess the "size" of the regions by Bayesian size, $b(x) = \pi(R_x \mid x)$ or confidence size (or cover) $C(\theta) = P(R_x \mid \theta)$. It then appears that the regions R_x are asymptotically of fixed confidence size if and only if

$$l(\theta) \propto I_{\mathscr{X}}^{\frac{1}{2}}(\theta) \propto I_{\mathscr{Y}}^{\frac{1}{2}}(\theta);$$

here $I_{\mathscr{X}}(\theta)$ is Fisher's information, and the "asymptotically" means that $\mathscr{X}$ and $\mathscr{Y}$ become infinite replicates of some base spaces $\mathscr{X}_0$ and $\mathscr{Y}_0$. The asymptotic behaviour of the tolerance regions R_x is independent of the prior density π, so that regions of fixed confidence size would seem to be inadmissible asymptotically unless the loss function K is generated by an $l(\theta)$ satisfying the above condition.

It is still reasonable to assess the size of the regions R_x by $C(\theta)$, θ unknown, as Mr Winsten has suggested; the Bayesian size $b(x)$ could be useful as an estimate of the confidence size. In fact $b(x) \to C(\theta)$ as $\mathscr{X}$ becomes an infinite replicate, and it may be reasonable to select the prior distribution π to make $b(x)$ a "good" estimate of $C(\theta)$ for finite $\mathscr{X}$.

Dr B. L. Welch: Theoretical distinctions are probably best brought out by the kind of discussion of specific problems which has been provided by the authors tonight. I should like, nevertheless, to make some points in the general theory of the single parameter, θ, which I believe to be relevant. It is well known that in large-sample Bayesian theory the choice of the prior distribution is not, within broad limits, a critical issue. Nor, if the appropriate confidence theory, based on the maximum-likelihood estimator, $\hat{\theta}$, is developed, do the numerical values of confidence points differ from those of Bayesian probability points. It is the present-day emphasis on small sample theory—possibly an over-emphasis—which opens up the breach between different approaches. It is of interest to note, therefore, that with a *particular* choice of prior distribution we can still secure concordance between Bayesian and confidence points at the further level of approximation where we take into account the first corrective terms in asymptotic theory, i.e. those terms whose influence is of order $n^{-\frac{1}{2}}$ in probability.

At this level the computation of confidence points in terms of $\hat{\theta}$ alone requires a knowledge of the third cumulant of $\hat{\theta}$. Alternatively, as has been described by Bartlett (1953), we may calculate confidence points on the basis of the distributional properties of $\partial L/\partial\theta$. Again the third cumulant is required. Although the results based on the distribution of $\hat{\theta}$ and those based on the distribution of $\partial L/\partial\theta$ are not identical they have the same probabilistic properties to order $n^{-\frac{1}{2}}$.

In Bayesian theory, on the other hand, it has been shown in a recent paper by H. W. Peers and myself (Welch and Peers, 1963) that there is one specific prior density function, $\omega(\theta)$, which has special properties in relation to confidence theory. The weight function is $\omega(\theta) = \sqrt{\{\kappa_2(\theta)\}}$ where $\kappa_2(\theta) = \text{var}\, \partial L/\partial\theta$, i.e. the weight function is equal to the standard deviation of $\partial L/\partial\theta$, or, near enough, is inversely proportional to the standard deviation

of $\hat{\theta}$. It was shown that Bayesian probability points calculated with this weight function can also be interpreted as confidence points. It can be proved furthermore that these confidence points, whilst not identical with those based purely on the distribution of $\hat{\theta}$, or on the distribution of $\partial L/\partial\theta$, have the same probabilistic properties to order $n^{-\frac{1}{2}}$.

In those situations where the variance of $\partial L/\partial\theta$ does not depend on θ to the order of $n^{-\frac{1}{2}}$ with which we are concerned $\omega(\theta)$ is constant and we have a uniform weighting. In general by an appropriate transformation $\phi = g(\theta)$ we can pass over to a parameter ϕ for which the variance of $\partial L/\partial\phi$ will be constant. The weighting will then be uniform on that scale. Furthermore, it is clear that this transformation is the familiar variance equalizing transformation which is commonly applied to the maximum-likelihood estimator $\hat{\theta}$. This, perhaps, is not unexpected.

For instance, in the case of binomial samples the variance stabilizing transformation is $y = \sin^{-1}\sqrt{(x/n)}$. Correspondingly we can write $\eta = \sin^{-1}\sqrt{p}$. Taking η to be uniformly distributed is the same as taking a weight function $\omega(p) \propto p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}$. One notes that in Mr Thatcher's paper he finds suggestions pointing to weight functions p^{-1} or $(1-p)^{-1}$ according to which of two problems he is considering. We may note that these two alternatives straddle the value $p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}$ suggested by general asymptotic theory. Perhaps too much should not be made of this since the latter theory assumes continuity for a variable which is strictly discrete. The step of the binomial variable expressed in terms of the binomial standard deviation is of order $n^{-\frac{1}{2}}$ and so discontinuity effects are, as Mr Thatcher has observed, not negligible in the present context. This may possibly be the whole explanation.

In general it may seem illogical to use frequency arguments to suggest prior distribution functions for insertion in Bayesian analyses. But historically this has often been done before—for instance in the case of the Student distribution. The sole object of these remarks is to point out that asymptotic theory taken beyond the usual large sample results can make suggestions of this nature. Whether Bayesians care to adopt them is another matter.

Professor G. A. BARNARD: It is not the custom at our Society to over-emphasize points of agreement and I will come straight to my disagreement, therefore, especially with Mr Aitchison. After so many speakers have drawn attention to the falsity of the antithesis between the Bayesian and the frequentist point of view it is perhaps not necessary for me to comment further. I cannot resist adding that one would hardly expect, after reading Mr Aitchison's paper, to find that one of the first men to advocate the application of Bayes's theorem, in circumstances like those contemplated here, was in fact Richard von Mises, who was in a sense the founder of the modern frequentist theory of probability. I know this false antithesis between "Bayesian" and "frequentist" methods does not originate with Mr Aitchison, but I think it would be very sad if it were given wider currency.

Another misapprehension that might arise also, and indeed apparently has arisen, is that the notion of introducing utilities into statistical decision problems is a specifically Bayesian notion. In fact it can be traced back to Gauss, but in more recent times it is primarily the work of Wald who was certainly never in any sense a subjective Bayesian.

The fact is, of course, that all the arguments which Mr Aitchison puts can be put into frequency terms, and, indeed, if I may refer to my own paper on "Sampling inspection and statistical decisions" (Barnard, 1954), especially pp. 163–165, I indicated there that in some respects the Bayesian argument can be better couched in terms of frequencies than in terms of probabilities.

In view of all that has been said and the hour, I must confine myself to making one further point. Consideration of the theory of tolerance intervals is valuable in connection with the Bayesian controversy, because there is another separation in the theory of tolerance intervals, that is, into the parametric and the non-parametric approach. It is well known that if we take the distribution to be continuous only, on being presented with the first

sample of n_1 observations, we can find limits associated with a small probability α such that unless an event of probability α occurs a specific proportion of a subsequent sample of n_2 will fall within given limits. This is without assumption about the form of the population. These limits will be expressed in terms of the percentiles of the sample. If we now assume that the population in question is (for example) normal, then we can find limits having similar properties, this time in terms of $\bar{x}$ and s. Whether we use the first set of limits or the second, that is, whether we use non-parametric or parametric limits, will depend on the *actual state of our knowledge*. If the assumption that the distribution is as nearly normal as no matter has empirical support of some kind, then we may use the parametric limits, but if we do not have such information then we shall be wise to use the non-parametric limits. It seems to me to be similar with regard to the non-Bayesian versus the Bayesian limits. We may *in fact know* the case in hand to be one of a series for which we can at least approximately guess the frequency function. If so, then we should use the Bayesian limits. But if not, then I think we should be deceiving our clients if we present them with Bayesian limits. If a professional body of statisticians were to sponsor the use of Bayesian tolerance limits in situations where nothing was known of the relative frequencies of the class of cases to which a given one was to be referred, it would be as if a body of professional civil engineers sanctioned a practice whereby if nothing was known of the subsoil upon which a building was to be erected the designer should assume it to be clay. What the civil engineers should do in such a case of course would be to make no such assumptions but instead to take steps to discover what the subsoil was. In the same way, if we are faced with a situation calling for Bayesian tolerance limits then it is our duty as statisticians to acquire the empirical knowledge necessary to support whatever assumptions we require to make about the prior distribution. Unfortunately, the treatment commonly given, and in this particular paper we have an instance of it, does not proceed in this way but instead proceeds by reference to what Raiffa and Schlaifer call conjugate prior distributions, or what I have called distributions closed under sampling. These distributions have mathematical elegance, but the fact that they have this gives us, to my mind, no reason whatever for supposing that nature will be so kind as to ensure that the frequencies met with in practice will follow such convenient forms. Just as Wald's minimax procedure implied the supposition that nature was unreasonably malevolent, I think the assumption of a convenient prior distribution assumes nature to be unreasonably benevolent.

I was interested to hear tonight that Professor Pearson made some empirical investigations to discover what nature's habits in these matters really were. It is very urgent that this kind of work should be pursued more vigorously. Mr Ford, at Imperial College, worked on this, too, and more recently Professor Hald in Denmark has done more work in connection with sampling inspection. There is a very unfortunate tendency amongst the fashionable Bayesian approach nowadays to ignore the importance of these empirical investigations.

May I conclude with one question for Mr Thatcher. He has not commented on the treatment of a related problem given by Fisher in terms of likelihood in his book *Statistical Methods and Scientific Inference* and I would be very grateful to hear his comments on that particular treatment.

Dr D. J. Bartholomew: I hope it is a sign of the times that papers have begun to appear which compare Bayesian and frequentist approaches to inference in the context of particular problems. The two theories can be mutually illuminating as I hope to show with particular reference to Mr Thatcher's paper. Before discussing my main point in detail it may help to clarify it by some general remarks.

Mr Aitchison's paper clearly demonstrates the mathematical advantages of Bayesian methods. Once the problem has been formulated the solution is a matter of mathematical routine. This is rightly contrasted with the ill-defined problem of choosing a pivotal function for frequentist methods. However, as Aitchison remarks; the price to be paid

for this simplification is the introduction of a prior distribution. To many statisticians this seems to require the abandonment of objectivity in favour of a theory built on the doubtful foundation of individual introspection. As Thatcher points out, the problem is particularly acute if our prior knowledge is vague or non-existent. However, it would be possible to retain the obvious advantages of the Bayesian approach without sacrificing objectivity if some formal way of representing ignorance could be found. It is well known that Jeffreys has suggested an invariance principle to achieve this end. (It is interesting to observe that his choice of prior distribution is the same as the weight function arrived at by Welch.) However, it is perhaps surprising that no one, to my knowledge, has suggested judging the suitability of a prior distribution solely by the frequency properties of the tests or estimates to which it leads. I hope to elaborate this principle elsewhere, but it has one consequence which is relevant to the discussion of Thatcher's problem. It implies that the choice of prior distribution and the choice of experiment are intimately linked. This means that the choice of a particular experiment expresses implicitly a certain prior belief about the unknown parameter. This is made explicit by finding what prior distribution gives results having the required frequentist properties. If there is no such distribution it seems plausible to conclude that the experiment proposed is not suitable. Conversely the principle requires that the Bayesian should select an experiment which, when combined with his prior distribution, leads to results which have the usual frequentist properties. The result of applying these general ideas to Thatcher's problem seems particularly interesting.

Consider it first from the frequentist point of view. It is important to recognize that Thatcher gives two distinct methods. One, giving upper limits, we shall call the U-method and the other, giving lower limits, we shall call the L-method. The U-method makes statements of the kind $\Pr\{a_2 > u_\alpha(a_1)\} \leqslant \alpha$ where α need not, of course, be small. Similarly for the L-method we have $\Pr\{a_2 < l_\alpha(a_1)\} \leqslant \alpha$. It is important to recognize that the two methods are different because it is not true in general that $u_{1-\alpha}(a_1) = l_\alpha(a_1)$ for all a_1. In fact it is easy to see from a diagram that $u_{1-\alpha}(a_1) > l_\alpha(a_1)$ for some values of a_1. Thatcher's two-sided prediction interval, with confidence coefficient at least $(1-2\alpha)$, is $\{u_\alpha(a_1), l_\alpha(a_1)\}$. It would be more natural to use either $\{u_\alpha(a_1), u_{1-\alpha}(a_1)\}$ or $\{l_{1-\alpha}(a_1), l_\alpha(a_1)\}$. The difficulty with both of these intervals is that they do not guarantee the required confidence coefficient although, as is clear from a diagram, it is possible to find a lower bound to the confidence. To do this we have to calculate the maximum probability outside the limits on any diagonal. An inspection of special cases suggests that this will seldom, if ever, be less than $(1-2\alpha)$. The U- and L-methods thus give confidence intervals. The intervals obtained by the U- and L-methods will never be longer than Thatcher's and are often shorter. Whichever method we use (U or L) the confidence coefficient is the same. The only difference between them is that the L-method gives shorter intervals if $a_1 < \frac{1}{2}n_1$ and longer intervals if $a_1 > \frac{1}{2}n_1$. In order to make a rational choice between the two methods we must have some prior information. All that is needed is some indication of whether p, and hence a_1/n_1, is more likely to be nearer 1 than 0 or vice versa. If this information is not available it can be argued that, in choosing U or L, we are acting as if we had it. If we are ignorant the simplest way to obtain the necessary information is to make a single observation before the experiment begins. If this were a success we would choose the U-method and if it were a failure the L-method.

Next consider the problem from the Bayesian point of view. Thatcher has shown that the U-method agrees with the Bayesian method if the prior density is proportional to $1/(1-p)$ and that the L-method agrees if the prior density is proportional to $1/p$. From the Bayesian point of view therefore the U-method would be appropriate if we had enough prior knowledge to believe that p was more likely to be near 1 than 0. The L-method would be used if we thought that p was more likely to be near 0 than 1. This is precisely the conclusion that we reached using frequentist arguments. We are now in a position to suggest a suitable prior distribution to represent complete ignorance about p. The minimum amount of information required to choose between the U- and L-methods is

that provided by one observation. We thus require a prior density which is converted into $1/(1-p)$ if the observation is a success and into $1/p$ if it is a failure. The only density satisfying this requirement is $1/p(1-p)$.

We have thus reached the following conclusion. The Bayesian and frequentist prediction limits agree under the following circumstances. The Bayesian must use the prior density $1/p(1-p)$ to represent complete ignorance. The frequentist must make one preliminary observation to decide whether to use the U- or L-limits.

This example illustrates the general point made earlier. The Bayesian formulation suggests that Thatcher's prediction interval can be improved upon since it cannot be derived from any prior distribution. It also makes explicit the fact that the U- and L-methods presuppose a very vague knowledge about p. This, in turn, indicates a way of modifying the experiment if we are completely ignorant. The frequentist approach, on the other hand, provides objective grounds for believing that the prior density proportional to $1/p(1-p)$ is a suitable representation of complete ignorance about p.

These conclusions have a bearing on Mr Aitchison's problem. Although the frequency statements about tolerance intervals may be tortuous they do provide some grounds for judging whether the prior distribution used is reasonable. It is thus reassuring to find that, in the simple examples considered, there is an equivalence between Bayes's and frequency methods.

Dr IRWIN: I should have thought that the suggestion that has been made of using the prior distribution $dp/p(1-p)$ would rather please the Bayesian because it would mean that $\log\{p/(1-p)\}$ would have a uniform distribution.

Mr H. E. BISHOP: Might there not be some advantage in not simply taking one preliminary observation to give a clue to the prior distribution? It might be better to take a number greater than one chosen to have some optimal property.

Dr BARTHOLOMEW: I have not investigated this point but my impression is that one observation would be the optimum number. This would leave the maximum number of observations for actually making the prediction.

Mr A. STUART: Like Professor Pearson and Professor Barnard, I lack strong belief in my mind as a mirror to nature, and so I am not given to persuading myself that merely convenient prior distributions really are the ones I should be using. Bayesians have sometimes justified such self-persuasions by pointing out that they are no worse than those involved in the common distributional assumptions made in parametric problems. This, it seems to me, is a stronger argument for abandoning the prevalent evil than for adding another to it, but Mr Aitchison dismisses the distribution-free approach in a sentence.

Bayesian statisticians depend upon a modern form of the pathetic fallacy—they expect the world to reflect the vibrations of their minds. I hope they will not model themselves too closely upon the hero of a satire who, when asked for a rule of invention, said: "Why, Sir, when I have anything to invent, I never trouble my head about it, as other men do." There is no escape from satire in the modern world, but this one, *The Rehearsal*, was published in 1672 by the second Duke of Buckingham. Perhaps with some pre-vision, he named his hero "Bayes".

Professor D. V. LINDLEY: Nearly all discussions of probability treat it as a function of a single argument. That is, with an event A of a suitable type (what is suitable depends on the school of probability) is associated a number called the probability of A and denoted by $p(A)$. This is false: probability is a function of *two* arguments; the event A being considered as above, and the conditions, B say, under which the consideration is taking place. That is, all probabilities are conditional and should be written $p(A \mid B)$. B is essential in the determination of the numerical value of the probability. Recognition

of this fact resolves many paradoxes. Thus if A is a hypothesis, it is often said that its probability, if it has one, is either 1 or 0 depending on whether it is true or false. This is only correct if B contains the knowledge of its truth or falsity: if B contains substantially less knowledge values other than 0 or 1 are perfectly sensible and permissible.

In a study of tolerance regions (and to use Mr Thatcher's notation) the practical circumstance is that one will have observed a_1 successes in n_1 trials and be desirous of making a probability statement about a_2, the number of successes in a future n_2 trials. In other words, the practical requirements demand consideration of $p(a_2; n_2 \mid a_1; n_1)$ in an obvious notation. Of what practical value then is the frequentist probability statement about $p(a_2 > u(a_1))$ when the real situation involves a knowledge of a_1? For this practical reason the frequentist approach to tolerance regions seems to me to be incorrect. Only a Bayesian can discuss the relevant conditional probability $p(a_2; n_2 \mid a_1; n_1)$.

A second point is that I have yet to meet a situation for which the problem of *restricted* tolerance regions is appropriate. Let us take, for example, the electrical supply system illustration of Mr Aitchison's. This can be formulated as a decision problem. The possible decisions are the pairs (r, s) of resistances and we can consider the utility of any pair given θ, the state of nature. The utility will involve the cost of the pair and also the more difficult considerations of the consequences of a drop in voltage below v. It may be objected that the latter are involved, but they are relevant to the problem and any avoidance of them by restriction to a value of q is an artificial device to avoid some hard thinking about the consequences. One is surely likely to obtain a better result by a deep consideration of the utility than by a mathematical restriction based on a practically confusing value q.

The following written contributions were received after the meeting:

Dr I. J. Good: I have read Mr Thatcher's paper and am happy to comment on it since I have for decades been interested in circumstances in which statisticians implicitly make use of initial distributions. For example, the fiducial argument occasionally implies an initial distribution (Lindley, 1958). But it can be used inconsistently. [If the probability densities for two experiments are $f(x, a)$ and $f(x, b)$ where

$$f(x, a) = \theta^2(x+a)\, e^{-x\theta}/(\theta+a) \quad (a>0,\ \theta>0,\ x\geqslant 0,\ a\neq b),$$

the final fiducial probability density for θ depends on the order in which the two experiments are reported!] "Pistimetric inference" for multinomial distributions (Roy, 1960) is clearly equivalent to the use of an initial density of the special divergent Dirichlet form Πp_i^{-1}, and then the final expectation of p_i is its maximum-likelihood value. For the binomial case one gets Haldane's distribution $p^{-1}(1-p)^{-1}$ (Jeffreys, 1961, p. 123). The estimation of a multinomial probability by Johnson (1932, pp. 421–423) as $(n_i+k)/\Sigma_i(n_i+tk)$ (where n_i is the ith frequency; $i = 1, 2, \ldots, t$) can be shown to be equivalent to the use of an initial density of the symmetrical Dirichlet form. Johnson's most controversial assumption was that the probability that the next category to be observed is the ith one depends on n_i and on the total sample size, Σn_i, but not otherwise on n_j for $j \neq i$. This assumption might be a reasonable approximation for small samples, but it is false, for example, in the multinomial sampling of species (Good, 1953; Good and Toulmin, 1956). In the binomial case, $t = 2$, Johnson's assumption is indisputable, but the beta distribution is not thereby established since his proof breaks down in this case, a fact that seems not yet to have been noticed.

A modern Bayesian who puts emphasis on the judgements of probability inequalities will regard as reasonable a high order of infinity of initial probability distributions, but will usually be prepared to limit his attention provisionally to a class with only a finite number of parameters. He might adopt Hardy's use of the beta distribution in 1889 (see Perks, 1947). This was perhaps the first use of Carnap's (1952) "continuum of inductive methods". The obvious generalization to the multinomial case is the class of Dirichlet distributions. A modern Bayesian might sometimes judge the use of only a finite number

of initial distributions, such as Mr Thatcher's p^{-1} and $(1-p)^{-1}$, to be adequate. But if the problem is one with logical symmetry between "successes" and "failures", then all reasonable initial distributions are symmetrical about $p = 1/2$. So I cannot readily accept the confidential predictions, but my reason would not be the speaker's since I *am* prepared to use more than one initial distribution at the same time. [Note that the average of p^{-1} and $(1-p)^{-1}$ is Haldane's $p^{-1}(1-p)^{-1}$. Perhaps this would lead to results *approximately* agreeing with the confidential predictions.]

In the multinomial case, the initial distribution proportional to Πp_i^{k-1} leads to the value $(k+1)/(tk+1)$ for the initial type II expectation of the "repeat-rate" Σp_i^2. You could select k by equating this initial expectation to your direct initial guess of the population parameter, or you could use upper and lower guesses and so upper and lower values for k. Or you might prefer to assume a type III distribution for k. Whatever method you used, you would *not* simply take $k = 0$, since the initial type II expected repeat-rate would then be 1. The pistimetrician, Roy, therefore implicitly uses an unreasonable initial distribution.

As a natural generalization, the Bayesian could select one of an m-parameter class of initial distributions by guessing initial values of m population parameters.

Theorem 1 is a special case of the following fact: If a parameter is repeatedly and independently selected from a superpopulation in accordance with some type II physical probability distribution (Good, 1957, p. 862), then a Bayesian who repeatedly uses any other assumption for the distribution will make probability estimates that will almost certainly not agree with the long-run frequencies. So a Bayesian must learn from experience and use his judgement rather than always adopting fixed rules. This is why some of us use subjective probability rather than credibility, but credibility is an ideal.

Mr Thatcher states that if $p = 1$ the Bayesian's predictions will be wrong every time, if he uses a continuous initial density. But they will not if he attaches non-zero probabilities to null hypotheses. If every trial in a large binomial sample is a success, the Bayesian would predict that *probably* $p = 1$. The speaker's apparent strictures concerning Bayesians apply only to out-moded varieties.

The confidence man, in the same circumstances, would merely put the upper bound of p at 1, in other words *he would say absolutely nothing* about the upper bound. Confidential methods need improving with the aid of a little utility philosophy. When the true value of a parameter is p, there is a utility loss $\Psi(S \mid p)$ in stating that the parameter lies in a set S, as compared with correctly stating the precise value of p. This loss decreases when the interval or set is made smaller, provided that p remains inside the interval: hence short confidence intervals are preferred to long ones. The function Ψ always lurks in the background, formless and unformalized.

Dr G. M. Jenkins: Mr Thatcher's result that the confidence limits for his problem correspond to certain Bayesian limits, although mathematically interesting, does not seem to me to be very relevant. Furthermore, the prior distributions which he obtains look rather curious.

I do not see that much is to be gained by judging Bayesian and likelihood methods by means of criteria which are relevant to decision theory. If one takes the view that inference is concerned with trying to extract as much information as possible from what might well be the one and only experiment one is able to perform, then the probability of being "correct" (whatever this may mean) in a hypothetically infinite sequence of experiments is irrelevant.

On the other hand, the Bayesian approach skates on thin ice if no objective prior distribution is available, since the answer is different for each prior distribution. In a recent paper (Barnard, Jenkins and Winsten, 1962, to be referred to as BJW) it was stated that the application of Bayes's theorem in any objective sense is equivalent to using the likelihood function from one experiment as the prior distribution for the next. For a series of experiments, this is equivalent to saying that the likelihoods are multiplied. As

emphasized in (BJW), an important problem in this area is the correct choice of a *metric* or *scale* on which to plot the likelihood function. One solution to this problem may be obtained by noting that if the log-likelihood function $\mathscr{L}(\theta) = \log L(\theta)$ is parabolic, i.e.

$$\mathscr{L}(\theta) = -(\theta - \hat{\theta})^2/2\sigma^2, \tag{1}$$

where $\hat{\theta}$ is the maximum-likelihood estimator, then the second derivative $-(d^2\mathscr{L}/d\theta^2)$ is constant and equal to $1/\sigma^2$. In the sampling theory approach it is the expected value of this which measures the information in θ. For a normal or parabolic likelihood this is constant and hence there is equal information for all values of θ. This suggests that a good choice of scale is to find a transformation $f(\theta)$ such that the likelihood is parabolic in f. From a Bayesian point of view it is this scale which seems the natural one in which to assume that the prior distribution is uniformly distributed.

It will not be possible in general to find a transformation $f(\theta)$ which makes the likelihood function exactly parabolic in f so that a less restrictive aim would be to make the second derivative independent of f in the neighbourhood of $\hat{f} = f(\hat{\theta})$, the maximum-likelihood estimator of f. Since

$$-\left(\frac{d^2\mathscr{L}}{df^2}\right)_{\hat{f}} = -\left(\frac{d\theta}{df}\right)^2_{\hat{f}} \left(\frac{d^2\mathscr{L}}{d\theta^2}\right)_{\hat{\theta}}, \tag{2}$$

it follows that the left-hand side is constant if

$$f = \int \sqrt{\left(\frac{-d^2\mathscr{L}}{d\theta^2}\right)_{\hat{\theta}}}\, d\theta. \tag{3}$$

If we take f to be uniformly distributed, then the prior distribution for θ is

$$\phi(\theta)\, d\theta = \sqrt{\left(\frac{-d^2\mathscr{L}}{d\theta^2}\right)}\, d\theta, \tag{4}$$

where again the second derivative is evaluated as a function of θ in the neighbourhood of $\hat{\theta}$.

For a binomial distribution

$$\left(\frac{d^2\mathscr{L}}{dp^2}\right)_{\hat{p}} = \frac{-n}{\hat{p}(1-\hat{p})}, \tag{5}$$

and hence $f = \sin^{-1}\sqrt{p}$ and (6)

$$\phi(p)\, dp = \frac{dp}{\sqrt{\{p(1-p)\}}}. \tag{7}$$

It is to be noted that the transformation (3) is mathematically equivalent to the variance-stabilizing transformation in sampling theory, but the logic underlying it is quite different. Other interesting prior distributions which emerge from this approach are given in the following table.

Parameter	*Scale*	*Prior distribution*
Normal variance	$f = \log \sigma$	$d\sigma/\sigma$
Exponential distribution	$f = \log \mu$	$d\mu/\mu$
Correlation coefficient	$f = \tanh^{-1} \rho$	$d\rho/(1-\rho^2)$
First-order autoregressive	$f = \sin^{-1} \alpha$	$d\alpha/\sqrt{(1-\alpha^2)}$

It is not suggested that a local transformation such as (3) will always be a good normalizing transformation but it does lead to intuitively sensible prior distributions. Professor H. E. Daniels has pointed out to me that for a binomial distribution, the logit transformation

$$f = \log\left(\frac{p}{1-p}\right) \tag{8}$$

would result in a likelihood function more nearly normal than the arc sin transformation.

This would correspond to a prior distribution of $dp/p(1-p)$. Analogous considerations arise in sampling theory where it is well known that a variance-stabilizing transformation is not necessarily the best normalizing transformation.

If one were pressed to give an answer to Mr Thatcher's problem, namely, based on a binomial sample with a_1 successes out of n_1 what are the "95 per cent limits" for the number of successes in a further sample of n_2, one could approximate the likelihood in the first sample by taking $\sin^{-1}\sqrt{p}$ to have a normal form with mean $\sin^{-1}\sqrt{\hat{p}}$ and variance $1/n_1$. If all we know is that a further sample of n_2 is to be taken then this will have the effect of reducing the variance to $1/(n_1+n_2)$, but our best estimate of the mean is still $\sin^{-1}\sqrt{\hat{p}}$. Hence "95 per cent limits" for $\sin^{-1}\sqrt{p}$ will be given approximately by

$$\sin^{-1}\sqrt{\hat{p}} \pm 1{\cdot}96/\sqrt{(n_1+n_2)}. \tag{9}$$

If several parameters are to be estimated then frequently the log-likelihood function factorizes as follows:

$$\mathscr{L}(\theta) = \mathscr{L}_1(\theta_1) + \mathscr{L}_2(\theta_2) + \ldots + \mathscr{L}_n(\theta_n). \tag{10}$$

Examples of this factorization have been given in (BJW).

Since

$$\frac{\partial^2 \mathscr{L}}{\partial \theta_i^2} = \frac{\partial^2 \mathscr{L}_i}{\partial \theta_i^2} \quad \text{and} \quad \frac{\partial^2 \mathscr{L}}{\partial \theta_i \, \partial \theta_j} = 0 \tag{11}$$

it follows that the scales for the n parameters may be chosen independently and the transformed likelihood behaves approximately like n independent normal likelihoods.

When the cross derivatives (11) are not zero, then the choice of scales requires the solution of certain partial differential equations. I have not yet been able to find a solution for a problem of this kind.

Mr Aitchison replied briefly at the meeting and subsequently in writing as follows:

I wish to express my thanks to Mr Winsten, Mr Kerridge and other speakers for their kind remarks about my paper and to all who contributed to a helpful discussion.

Before I consider the more controversial aspects of the discussion I wish to comment on two technical points. Dr Hartigan's simplification when the cost function K in the value function (39) is a positive measure is extremely elegant. Since cost will often only be meaningful for intervals of the form $(-\infty, r)$ or $(0, r)$, or their higher-dimensional counterparts, and since cost will increase with r the positive-measure assumption is realistic and his development, together with its accompanying concept of Bayesian size, could be of decided practical value. In my paper I concentrated too closely on cases where the informative experiment consists of replicates of the basic or prediction experiment to notice the advantages, indicated by Dr Hartigan, of completely divorcing the observation and prediction spaces. I adopt this approach now to make my second technical point—not commented on in the discussion—namely, the connection between the two papers. In this more general setting the frequentist and Bayesian problems, whose equivalence or lack of it are Mr Thatcher's concern and whose probabilistic formulations are his relations (3) and (13), are easily shown to be restricted tolerance region problems. If x denotes the observation on the informative binomial experiment, then (3) and (13) seek a region R_x which is a set of integers $\{0, 1, \ldots, u(x)\}$. If y denotes a typical outcome of the binomial experiment for which prediction is required then we can arrive at the two probabilistic statements by using the restricted value function

$$V(R_x, y) = \begin{cases} 1 & (y \in R_x), \\ 0 & (y \notin R_x), \end{cases}$$

and setting the corresponding expected utility equal to $1-\alpha$, expectation for the frequentist being with respect to the binomial density $p(x \mid \theta)$ and for the Bayesian with respect to the posterior density $\pi(\theta \mid x)$, where my θ corresponds to Mr Thatcher's p.

Mr Winsten's suggestion that, for any proposed region R, the graph of the cover or payoff function $P(R|\cdot)$ should be investigated to highlight the awkward values of θ for that R is a good one, easily missed in the struggle to reach a decision. But it is necessary to decide eventually on some R and I am not at all clear how Mr Winsten proposes to decide. It is unhappily a common feature of such decision problems that no single reasonable R is outstandingly successful for all θ or even all probable θ. In his apparently informal use of his "ancillary evidence" about θ I suspect that he may be eventually indistinguishable from a Bayesian.

It is a common frequentist tactic to make the disarming claim that there is no antithesis between frequentism and Bayesianism and that the frequentist is prepared to admit Bayesian methods, even to say that they are the appropriate tool, in their proper place. Both Mr Kerridge and Professor Barnard use this tactic in their own individual ways. Of course frequentist von Mises advocated the use of Bayes's theorem and of course frequentists developed decision theory. My concern was the present state of tolerance region theory, the inability of users to comprehend the frequentist formulation, and the lack of evidence of any serious attempt among statisticians (Professor Barnard and Mr Kerridge excepted) to practise what they preach and actually ask their clients if they have any prior information about θ. I was disappointed, though not surprised, to find no frequentist rise to the defence of his formulation by way of the principle, so repugnant to the Bayesian, of imbedding a decision problem in a hypothetical long run of such problems with its use of irrelevant probabilities such as $P^{(n)}(\cdot|\theta)$, a point well stressed by Professor Lindley. A re-reading of my paper has made me realize that I was still too frequentist in outlook when I wrote it. While trying to provide alternatives to the frequentist probabilistic setting I did not free myself sufficiently from other aspects of that theory such as the concept of quality q, which I would now happily abandon in favour of more formal Bayesian decision theory. Abandonment of q does not, however, imply that I agree with Professor Lindley that the restricted problem as I defined it is not a real problem. Mr Maver, in his very helpful contribution as a user, gives examples which I hope provide the necessary existence theorem.

I do not now attach much importance to attempts to establish equivalence of Bayesian and frequentists results, as considered in Section 4 of my paper, and as studied by Mr Thatcher in his paper, and Mr Kerridge and Dr Welch in the discussion. To a Bayesian there seems to be no especial merit in showing that one's own formulation can lead—under certain circumstances—to results equivalent to those of another formulation which one regards as absurd. To judge a Bayesian solution solely by its frequency properties, as suggested by Mr Kerridge and Dr Bartholomew, seems to me equally open to question (a point also made by Dr Jenkins) though I had better not prejudge Dr Bartholomew's promised results. Mr Kerridge has, however, propounded his assessment of Bayesianism in forthright frequentist terms but his concept of "the right answer" completely eludes me. How, if we are faced with a new type of decision problem are we to decide, at the *relevant* time of making the decision, whether we are making the right decision? Who is to be the arbiter of right or wrong? I would be as reluctant to accept Mr Kerridge's arbitration as he would be to accept mine. Again, since I cannot find the lines between which he was reading, I see no need to answer his charge that I am trying to tell engineers what they ought to believe. His experience of engineers is obviously slight if he imagines I have such ambitions.

Many speakers point to a lack of objectivity in the use of prior densities and in particular Professor Barnard and Mr Stuart conjure up a picture of the Bayesian as an ivory-towered philosophizing fool. It seems to me that the fallacy of all such arguments is the naïve view that decision-making is, or can be made to be, an objective occupation. On the contrary, it requires judgement all along the line. The choice of a loss or utility function is clearly a matter for informed assessment, for who can categorically say that conditions and consequently losses will not have altered by the day of reckoning. The choice of a parametric or a distribution-free model is another matter of nice judgement.

(My "dismissal" of distribution-free regions, which offends Mr Stuart, was on the grounds that they are open to the same anti-frequentist attack as parametric regions and, of course, provide no direct comparisons with the Bayesian formulation.) Professor Pearson rightly points out that one's system of inference is a purely personal attribution. To him the use of prior densities often seems dishonest. To me the principle of imbedding decision problems calls for as much introspection as the search for prior information about θ, especially as I believe, with Mr Winsten, that such information exists for those who seek it. Professor Barnard is probably correct when he complains that Bayesian writing, including my own, tends to be couched in terms which give the impression of an unwillingness to acquire empirical knowledge about θ, and I hope we heed his warning in future. There do remain, however, situations where the only way to acquire information about θ is by performing the informative experiment indexed by θ. For these I do not see that I deceive my client if I ask for *his* views about θ, show him, and allow him to adjust, his prior density function—the use of conjugate prior densities is a matter of convenience and not essential to the Bayesian; work on the mathematical expression of ignorance, such as that of Dr Jenkins, is very relevant here. How often does the frequentist deceive his client by not telling him that, for the purposes of solution, his particular decision problem has been imbedded in a sequence of such decision problems. The strength of the Bayesian approach to any problem lies in directing attention to those aspects of the decision problem which are relevant and in displaying clearly how, and on what assumptions, the decision is being taken. It might even be argued that by highlighting the subjective elements in decision-making the Bayesian makes the process more objective.

Mr THATCHER replied briefly at the meeting and subsequently in writing as follows:

I should first like to thank Professor Pearson for his fascinating contribution, and Dr Good for his masterly conspectus of the neo-Bayesian position.

So much of the discussion seemed to be in tacit agreement with a "Bayesian-frequentist" approach that, like Mr Kerridge, I find it unexpectedly difficult to discover points of controversy. Fundamental attitudes have had a thorough airing. As a change, I should like to consider certain mathematical points raised by Drs Bartholomew, Jenkins and Welch, before returning to a final question of principle.

The results in Sections 2 and 7 of my paper were deliberately confined to one-sided confidence limits and their associated two-sided central limits (i.e. such that the frequency of wrong predictions does not exceed α at either limit). Dr Bartholomew introduces the much more tricky subject of non-central limits (i.e. such that the frequency of wrong predictions does not exceed 2α at the upper and lower limits taken together, but may exceed α at one or the other). The distinction is important. Dr Bartholomew describes my derivation of central limits as a mixture of two different methods, and suggests that a "rational choice" between them, based on a preliminary observation, should be used to obtain non-central limits. I am not quite clear why a preliminary observation followed by a sample of size n_1 should be very different from a sample of size n_1+1; but, be that as it may, there are other ways of finding non-central limits. For example, several methods are quoted by Blyth and Hutchinson (1960) for the case when n_2/n_1 is infinite.

Dr Bartholomew's arguments lead him to the conclusion that, in the problem of two binomial samples, a Bayesian who uses the prior distribution $p^{-1}(1-p)^{-1}\,dp$ will make predictions acceptable to a frequentist. But with this prior distribution, whenever the first sample consists entirely of successes ($a_1 = n_1$), the Bayesian probability distribution for a_2 is entirely concentrated at $a_2 = n_2$. Consequently the Bayesian, whether he uses central or non-central limits, will be bound to predict that the next sample will contain no failures at all. So even if the predictions are right whenever $a_1 < n_1$, they will still be wrong whenever $a_1 = n_1$, $a_2 < n_2$; and by the method of Section 5.2, case B, we can find an example in which the frequency of these wrong predictions will exceed 2α. This proves that Dr Bartholomew's conclusion is false.

The same disadvantage applies to the prior distribution $p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}\,dp$. Drs Jenkins and Welch show that it has desirable asymptotic properties; but, nevertheless, predictions based on it do not satisfy the frequency criterion. More generally, Theorem 1 shows that it is impossible for a Bayesian to find a prior distribution which will satisfy a frequentist.

My aim in seeking one was, of course, the same as Dr Bartholomew's; but Theorem 1 shows that one prior distribution is not enough. It is therefore interesting to learn that so eminent a Bayesian as Dr Good is prepared to use more than one prior distribution at the same time; though he objects to the particular pair $p^{-1}\,dp$ and $(1-p)^{-1}\,dp$, on the grounds that they do not treat successes and failures symmetrically. However, this objection is overcome by the "strategy" in Section 7.7 of the paper. I was sorry that none of the Bayesian contributors made comments on this point. Subjective choices from an infinity of prior distributions may be all very well when there is *some* ground for making the choice; but when there is no prior information whatever, can one really improve on a randomized strategy?

References in the Discussion

Barnard, G. A. (1954), "Sampling inspection and statistical decisions", *J. R. statist. Soc.* B, **16**, 151–165.

—— Jenkins, G. M. and Winsten, C. B. (1962), "Likelihood inference and time series", *J. R. statist. Soc.* A, **125**, 321–372.

Bartlett, M. S. (1953), "Approximate confidence intervals", *Biometrika*, **40**, 12–19.

Carnap, R. (1952), *The Continuum of Inductive Methods*. Chicago University Press.

Edgeworth, F. Y. (1884), "The philosophy of chance", *Mind*, **9**, 223–235.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics", *Phil. Trans. Roy. Soc.* A, **222**, 309–368.

Good, I. J. (1953), "The population frequencies of species and the estimation of population parameters", *Biometrika*, **40**, 237–264.

—— (1957), "Saddle-point methods for the multinomial distribution", *Ann. math. Statist.*, **28**, 861–881.

—— (1960), Contribution to the Discussion on Mr. Beale's paper, *J. R. statist. Soc.* B, **22**, 79–82.

—— and Toulmin, G. H. (1956), "The number of new species, and the increase in population coverage, when a sample is increased", *Biometrika*, **43**, 45–63.

Johnson, W. E. (1932), "Probability: the deductive and inductive problems", *Mind*, **41**, 409–423.

Kerridge, D. (1963), "Bounds for the frequency of misleading Bayes inferences", *Ann. math. Statist.*, **34**, 1109–1110.

Lindley, D. V. (1958), "Fiducial distributions and Bayes' theorem", *J. R. statist. Soc.* B, **20**, 102–107.

Pearson, E. S. (1925), "Bayes' theorem, examined in the light of experimental sampling", *Biometrika*, **17**, 388–442.

Pearson, K. (1920), "The fundamental problem of practical statistics", *Biometrika*, **13**, 1–16.

Perks, W. (1947), "Some observations on inverse probability including a new indifference rule", *J. Inst. Actuar.*, **73**, 285–312.

Roy, A. D. (1960), "Some notes on pistimetric inference", *J. R. statist. Soc.* B, **22**, 338–347.

7 PREDICTIVE DENSITY FUNCTIONS

From the form of the expected utility $W(R,x)$ in §6.2 it is clear that

$$p(y|x) = \int_{\theta} p(y|\theta)p(\theta|x)d\theta$$

must play a central role in predictive problems. In relation to e and f this *predictive density function* sums up our present view, based on e and $p(\theta)$, of the relative probabilities of the outcomes of f. Since any analysis concerning f must be undertaken at the present it is the relevant inference tool for predictive purposes.

The idea was not new, the 'rule of succession' of Laplace, concerned with the probability of success at the (n+1)th binomial trial given n successes in the preceding n trials, being an early particular example. Jeffreys (1961, p.143), Geisser (1964) and Guttman and Tiao (1964) had examined situations involving predictive density functions associated with normal distributions, with Guttman and Tiao also considering two-parameter negative exponential distributions. All of these used conventional vague prior distributions. The aim of Aitchison and Sculthorpe (9:1965) was to emphasise the key role that such predictive density functions play in predictive problems, and a first step was the derivation of predictive distributions for the standard univariate models - binomial, Poisson, gamma, normal, normal linear regression - on the basis of proper conjugate prior distributions.

In addition to clarifying the structure of predictive problems Aitchison and Sculthorpe (9:1965) are concerned with the provision of a viable means of resolving some engineering design problems.

For this purpose they provide a catalogue of expected utilities for various simple utility functions so that a user, by expressing his more complex utility function in terms of these simpler functions, has an immediate evaluation of expected utility for decision purposes.

The framework for statistical prediction contained in Aitchison and Sculthorpe (9:1965) allows them to bring within one general framework previous work on Bayesian and frequentist tolerance regions, to classify these regions in terms of simple specifications of the utility U or value V function and to identify situations in which such tolerance regions may be appropriate. Aitchison (10:1966) later shows that a linear-utility tolerance interval, say with $R = (-\infty, r)$ and

$$V(R,y) = \begin{cases} y-r & (y \leq r), \\ \lambda(r-y) & (y > r), \end{cases}$$

is identical to an expected-cover tolerance interval with cover $\lambda/(1+\lambda)$. This result thus provides users of expected cover tolerance intervals with a decision-theoretic interpretation for their choice.

AITCHISON, J. and SCULTHORPE, Diane (1965)

Some problems of statistical prediction

Reprinted from *Biometrika* 52, 469-83

Biometrika (1965), **52**, 3 *and* 4, *p.* 469
With 2 text-figures
Printed in Great Britain

Some problems of statistical prediction

BY J. AITCHISON AND DIANE SCULTHORPE
University of Liverpool

SUMMARY

A general framework is introduced for the study of inference and decision predictions about the outcome of a future experiment from the data of an independent informative experiment. This allows a simple classification of prediction problems, and shows the place of standard inference predictions within the framework. A Bayesian approach to decision prediction is then presented and techniques appropriate to a variety of realistic utility functions are developed. Finally, some prediction problems associated with classes of experiments are considered.

1. INTRODUCTION

Statistical prediction is the use of the data from an informative experiment E to make some statement about the outcome of a future experiment F. The prediction statements commonly treated in the literature are of inference type, in which the purpose is to give some indication of the likely outcome of F, or to suggest some subset of possible outcomes in which the actual outcome of F is likely to fall. There are also, however, prediction problems of a decision type, for which the decision space consists of subsets of the outcome space of F, and where the prediction is related in a much more precise way to some specific purpose. Our own interest in the subject has arisen from decision problems in the supply of hospital engineering services (e.g. oxygen, gas, conditioned air, suction, etc.). In a simple version the supply system may be supposed to function at a series of independent operations, at each of which a constant quantity r (e.g. number of outlets) of the commodity is available for supply. At each operation of the system some variable quantity y is demanded; this may be below or above r. If $y > r$ the system has failed fully to meet demand and if $y < r$ the system has oversupplied. The extent to which fixing the supply at r is satisfactory depends on the relative demerits of failing to meet demand and of oversupplying, and on the variation in y. Here we can suppose F to be the observation of a free demand, unrestricted by the limited supply. The informative experiment E may consist of demands $x_1, \ldots, x_n$ on an existing similar system which has been overdesigned, so that E consists of n replicates of F. If the existing system is not overdesigned but supplies r_1, say, at each operation then E may be regarded as n replicates of F, with observations truncated or censored at r_1; this case can be treated only by asymptotic methods and we shall not consider it here.

Our purpose in this paper is first to suggest a clear and flexible framework within which such inference and decision prediction can be discussed, to indicate briefly how existing inference procedures fall within this framework, and then to develop the model towards specific decision prediction procedures. We do this for the case in which E and F are independent experiments; thus we do not consider the situation where the outcome of E is a part realization of a stochastic process and F is the continuation of the process.

The four ingredients of this type of prediction problem are as follows.

(i) *The future experiment.* There is a future experiment F for which a prediction of some

sort is required. We suppose that F has outcome space Y (with typical outcome $y \in Y$) and event space $\mathscr{Y}$, and that the possible probabilistic descriptions of F form the class of density functions

$$\{p_F(\cdot|\theta): \theta \in \Theta\},$$

where the parameter space Θ is some part of a finite-dimensional Euclidean space. We denote by $P_F(\cdot|\theta)$ the probability measure on $\mathscr{Y}$ corresponding to the density function $p_F(\cdot|\theta)$ on Y. The true value of the parameter θ is not precisely known.

(ii) *The informative experiment.* An informative experiment E has been performed. This experiment we suppose to have outcome space X (with typical outcome x) and event space $\mathscr{X}$, and to be described by one of the class of density functions

$$\{p_E(\cdot|\theta): \theta \in \Theta\};$$

corresponding probability measures on $\mathscr{X}$ are denoted by $P_E(\cdot|\theta)$. The choice here of θ as indexing parameter is quite deliberate for we suppose that the true describing densities of E and F have the same true parameter value as index. It is through this connexion between E and F that E provides information about F. (The full description of E will often contain nuisance parameters, but no confusion should arise if we omit specific mention of them in the indexing of the class of density functions for E.) Although E and F are connected by θ we shall assume that, *for given* θ, they are statistically independent. We then denote the probability measure associated with the compound experiment (E, F) by $P_{EF}(\cdot|\theta)$; it will, of course, be the product probability measure associated with $P_E(\cdot|\theta)$ and $P_F(\cdot|\theta)$.

(iii) *The inference or decision space.* The characteristic of all prediction problems is that the stated inference or decision is a subset (possibly a point) in the outcome space Y of F. We can therefore conveniently take as inference or decision space the event space $\mathscr{Y}$ of F. If we observe $x \in X$ in the performance of E we take some region $R = \delta(x) \in \mathscr{Y}$ as our prediction region. The function $\delta: x \to \delta(x)$ is the inference or decision function. A special feature of such prediction problems is that there is a *natural* probability measure $P_F(\cdot|\theta)$ associated with the decision space; this feature is important in subsequent analysis.

To distinguish between the function δ and its value $\delta(x)$ for a particular outcome x of E we shall use the terms *predictor* and *prediction*, respectively.

(iv) *Evaluation of a predictor.* To choose between alternative prediction procedures it is necessary to have some means of evaluating a procedure. Many such evaluations are possible and the one chosen should always accord with real assessments. The most direct assessment (and one which appeals to the practical man) of the merits of a choice R for a prediction region is to consider its effectiveness in relation to each possible outcome y of F. For any chosen R and observed y it should be possible to assign some realistic utility or, since we are going to have two aspects of utility, a y-utility or *value* $V(R, y)$.

We now explore the possible methods of analysis from Bayesian and frequentist viewpoints to see how so-called standard techniques fit into the general framework and for what type of prediction they are formulated, and to set the scene for further development of the problem.

Two cases may be distinguished at this stage.

Case 1. A prediction is required for only one performance of F.

Case 2. It is envisaged that a series of replicates of F is to be conducted and that the prediction region R is to be used for each replicate. An example of such a situation is the demand and supply one already mentioned, where a replicate corresponds to an operation

of the system and the constancy of R to the fact that supply is to be the same at each operation. In such a situation we shall assume, except in § 6·1, that utilities are additive over replicates; this should hold true in many practical situations.

Faced with case 1 a Bayesian would proceed to obtain $\pi(y|x)$, the posterior distribution of y given x, probably through intermediate attention to the nuisance parameter θ; from a prior density $\pi(\theta)$ on Θ the posterior density $\pi(\theta|x)$ is obtained in the usual way and this is converted into $\pi(y|x)$ through the relation

$$\pi(y|x) = \int_\Theta p_F(y|\theta)\,\pi(\theta|x)\,d\theta. \tag{1}$$

(Integration should be replaced by summation when discrete spaces are involved.) The Bayesian then concerns himself with the expected utility with respect to this density, namely

$$H(R,x) = \int_Y V(R,y)\,\pi(y|x)\,dy. \tag{2}$$

We shall see in succeeding sections exactly how the Bayesian makes use of the H function.

Since, from (1) and (2),

$$H(R,x) = \int_\Theta \left\{ \int_Y V(R,y)\,p_F(y|\theta)\,dy \right\} \pi(\theta|x)\,d\theta \tag{3}$$

we see, by defining a θ-utility $U(R,\theta)$ for R—or briefly a utility since we have decided to use the term value for y-utility—as

$$U(R,\theta) = \int_Y V(R,y)\,p_F(y|\theta)\,dy, \tag{4}$$

that the problem may be regarded from another viewpoint, namely as a Bayesian analysis in terms of the utility function U and the posterior density $\pi(\theta|x)$. Indeed in some inference problems, such as tolerance region predictions, the utility function is the starting point and expression is given to the Bayesian method by the use of

$$H(R,x) = \int_\Theta U(R,\theta)\,\pi(\theta|x)\,d\theta. \tag{5}$$

Relation (4) shows the interconnexion of the utility and value functions. It is clear that in practice each value function leads to a utility function, but it is possible to consider utility functions which do not arise from value functions. The two approaches—through value and utility functions—are mathematically speaking only different techniques of evaluating the double integral (3). If the θ-integration is performed first we use (1) and (2); if the y-integration is carried out first we use (4) and (5). Which we use can therefore be a matter of choice; it may be that one method is more simple mathematically than the other.

Case 2 leads to the same $H(R,x)$ and here the more natural road is through (4) and (5). For, if the region R is going to be used repeatedly in a sequence of independent performances of F and the value is $V(R,y)$ when y is observed, then since utilities are assumed additive over independent future experiments, $U(R,\theta)$ corresponds to the average value per replicate, and so provides a measure of the effectiveness of the prediction region. Note that the additive property of utilities is necessary only in case 2 to provide this appropriate frequentist interpretation of $U(R,\theta)$.

Although the value function V is an appealing one to the practical man it has some considerable conceptual difficulties for the frequentist. For instance, in case 1 it is natural for

him, in accord with frequentist decision theory, to introduce his predictor δ and to take the expectation of $V\{\delta(\cdot), y\}$ with respect to the informative density $p_E(x|\theta)$. The resulting expectation depends on (θ, y), the unknown state of nature; the presence of both θ and y makes the usual awkward feature of frequentist theory, namely the difficulty of finding pivotal statistics, even more embarrassing. The dependence of y on θ through $p_F(y|\theta)$ is waiting to be used but there is no obvious frequentist way to introduce it. For case 2, the argument which takes the value function to the utility function (see equation (4)) by way of a sequence of performances of F, allows the frequentist to proceed in his usual way by basing his considerations on

$$G(\delta, \theta) = \int_X U\{\delta(x), \theta\} p_E(x|\theta)\, dx. \tag{6}$$

We shall see in § 4, however, that this can lead to an inconsistency in interpretation, and there is little doubt that the frequentist is on the safest logical ground—though farther removed from practical considerations—when he confines himself to utility functions.

The introduction of the value function does, however, have some advantages for us in that it allows a convenient means of cataloguing problems so that frequentist and Bayesian counterparts are clearly displayed.

Fig. 1 displays the steps leading to the G and H functions which form the basis of frequentist and Bayesian prediction analysis. Where a density appears beside an arrowed line this means 'integrate with respect to the density over the appropriate space'.

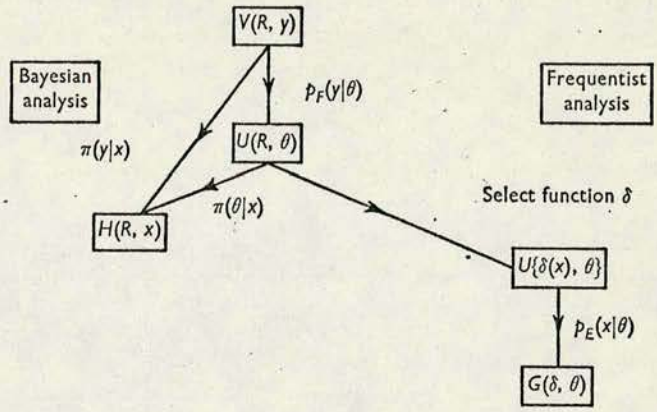

Fig. 1. Derivation of the basic criteria of Bayesian and frequentist prediction analysis.

2. CLASSIFICATION OF PREDICTION PROBLEMS

We have already divided problems into two categories by the Bayesian and frequentist approaches. While this is often for many statisticians a division on philosophical, and almost emotional, grounds we shall attempt not to enter into controversy here but merely state techniques and suggest under what conditions we think they are certainly applicable. For easy future reference we therefore first set out these Bayesian and frequentist conditions. We can then concentrate attention on the other four categories of classification.

Conditions for Bayesian analysis. All the Bayesian analyses of this paper are applicable when one or other of the following conditions obtains.

B 1. There is available, previous to E, some information on θ which can be described in terms of a prior density $\pi(\theta)$ on Θ.

B 2. There is no such prior information available but it is possible to describe 'ignorance' or 'vague previous information' in terms of a (possibly improper) density $\pi(\theta)$ on Θ; see, for example, Lindley (1965, pp. 13–18).

B 3. No generally acceptable $\pi(\theta)$ is available, but it is agreed that a useful method of drawing an inference or reaching a decision is to show the variation in R caused by choice of different $\pi(\theta)$.

Conditions for frequentist analysis. We shall refer to the following frequentist conditions when we judge that an analysis is applicable under one or other of them.

F 1. A series of E's is envisaged, each followed by a single F for which a prediction is required. A series of case 1 predictions is thus required.

F 2. Again a series of E's is envisaged. Following each E, however, a prediction is required for a whole series of F's. Thus we have in this case a series of case 2 predictions.

Conditions F 1 and F 2 usually allow a direct correspondence between long-run effects, e.g. relative frequencies and average utilities, and concepts of the mathematical model, e.g. probabilities and expected utilities. When there is no reality in the series of similar prediction problems envisaged in F 1 and F 2 then recourse has to be made to some principle of imbedding the problem in a series of dissimilar inference or decision problems; see, for example, a recent advocacy of such a principle, by Neyman (1964, pp. 927–32). We shall find it convenient to label this as a third frequentist condition.

F 3. Acceptance of the principle of imbedding.

(1) *The relationship of E and F*

Two main cases may be distinguished; these both arise frequently in practice.

(*a*) E is n replicates of F;

(*b*) F is a future experiment carried out at some value z of an independent variable, so that the future experiment is better described by F_z. E is then a *regression* experiment, i.e. it consists of independent experiments $F_{z_1}, \ldots, F_{z_n}$. We repeat that so long as E and F are indexed by the same parameter, prediction is possible; thus the case where E is n replicates of F, truncated at t, i.e. where

$$p_E(x|\theta) = p_F(x_1|\theta)\ldots p_F(x_n|\theta)/[1-P_F\{(-\infty, t)|\theta\}]^n,$$

can be readily treated by asymptotic methods. We shall, however, concentrate on the important cases (*a*) and (*b*) here.

(2) *The form of density*

We shall consider in this paper the standard distributions—binomial, Poisson, gamma and normal. It is possible to include cases (*a*) and (*b*) above within one framework and this we shall do here. The four specifications are as follows.

D 1. *Binomial.* E is Bi(k, θ), i.e. it consists of k binomial trials, each with probability θ of success; F is Bi(l, θ). Note that, by sufficiency arguments, this covers both cases (*a*) and (*b*). For (*a*), the total number x of successes in the n replicates is a sufficient statistic for θ and is a Bi(nk, θ) random variable. Hence take $k = nl$. For case (*b*), F_z is Bi(z, θ) and x is the total number of successes in $F_{z_1}, \ldots, F_{z_n}$; then x is sufficient for θ and is a Bi$(z_1 + \ldots + z_n, \theta)$ random variable. Such a regression situation may arise if the trials consist of processing a number of objects, the number of objects having arisen from some chance mechanism, e.g. a Poisson process. Note also that case (*a*) may be of interest in its own right; see, for example, Thatcher (1964).

D 2. *Poisson.* E is Po($k\theta$), i.e. E records a Poisson count with mean parameter $k\theta$ and F is Po($l\theta$). Remarks similar to those for the binomial case apply here.

D 3. *Gamma.* E is Ga(k, θ), with k known, i.e.

$$p_E(x|\theta) = \theta^k x^{k-1} e^{-x\theta}/\Gamma(k),$$

and F is Ga(l, θ). Similar remarks apply here also. For example, when E is n replicates of F the sum x of the outcomes of the n replicates is sufficient for θ and is Ga(nk, θ). Hence take $k = nl$ to obtain the above specification.

D 4. *Normal.* F is N(μ, σ^2) and E produces a joint sufficient estimate (m, s^2) for $\theta = (\mu, \sigma^2)$. We shall suppose that m, s^2 are independently distributed, and that m is N($\mu, h\sigma^2$) and $\nu s^2/\sigma^2$ is $\chi^2(\nu)$, where h and ν are known constants. Case (a) is clearly included in this. So also is the regression situation, where F_z is described by a N($\alpha + \beta z, \sigma^2$) density. For, if $\bar{x}$ and $\bar{z}$ have their usual meanings and

$$\hat{\beta} = \Sigma(x-\bar{x})(z-\bar{z})/\Sigma(z-\bar{z})^2, \quad s^2 = \Sigma\{x-\bar{x}-\hat{\beta}(z-\bar{z})\}^2/(n-2),$$

then, for given $\Sigma(z-\bar{z})^2$, $\{\bar{x}+\hat{\beta}(z-\bar{z}), s^2\}$ are jointly sufficient for $\mu = \alpha+\beta z$ and σ^2; also $\bar{x}+\hat{\beta}(z-\bar{z})$ is distributed as N$[\mu, \{1/n + (z-\bar{z})^2/\Sigma(z-\bar{z})^2\}\sigma^2]$, independently of $(n-2)s^2/\sigma^2$ which is $\chi^2(n-2)$. We consider the case of real z here for simplicity; there is no difficulty in extending the results to the case of vector z.

(3) *The utility specification*

In some problems, especially of the frequentist tolerance region type, the specification of utility is directly in terms of the utility function U, and there is no corresponding V, whereas in most practical situations it seems natural to specify in terms of a value function V. Thus it is first necessary to distinguish whether a U or V specification is involved. Then, of course the specific form of U or V involves a classification. We shall see later how a useful library of such specifications can be built up.

(4) *Inference or decision prediction*

The distinction we adopt here between a decision and an inference problem is that the former uses value and utility functions which lead to a genuine problem of finding a predictor δ which maximizes $G(\delta, \theta)$ for all θ in the frequentist context or a prediction R which maximizes $H(R, x)$ in the Bayesian context; whereas, inference problems, because of their less specific purpose, are based on very simple value and utility functions, and for these maximization of $G(\cdot, \theta)$ or $H(\cdot, x)$ leads to the trivial and useless statement that the appropriate prediction region is Y regardless of what x is observed. Inference predictors and predictions are therefore sought which yield some specified value, say q, below the maximum attained by Y. A frequentist inference prediction problem thus requires the finding of a predictor δ such that

$$G(\delta, \theta) = q \tag{7}$$

for all θ. The Bayesian counterpart is that of finding a prediction R (it is not necessary to find the predictor) such that

$$H(R, x) = q. \tag{8}$$

For discrete distributions such as D 1 and D 2 the equality sign has to be replaced by $\geqslant$.

Later, in § 5, we shall consider another class of problem where a prediction is required for a whole class of future experiments and not just a single F.

3. Preliminary distribution results

For Bayesian analyses it is extremely convenient to separate the construction of $\pi(y|x)$ and the specification of $V(R,y)$. Here we collect in a convenient form for further reference the appropriate results for the four standard density specifications D 1–4. Although a posterior distribution on Y is the object of this inference side of the analysis it is more natural to commence the argument through a prior density $\pi(\theta)$ on Θ. This yields, through the information x from E, a posterior density $\pi(\theta|x)$ on Θ, given x. Since the transition from $\pi(\theta)$ to $\pi(\theta|x)$ is now well catalogued (see, for example, Raiffa & Schlaifer, 1961) we shall suppose, to avoid over-elaborate notation, that in all our cases the stage $\pi(\theta|x)$ has been reached and that $\pi(\theta|x)$ has one of the following forms in the standard problems. The parameters of these distributions will, of course, depend on x.

D 1. *Binomial*
$$\pi(\theta|x) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b) \quad (0 \leqslant \theta \leqslant 1). \tag{9}$$

D 2, D 3. *Poisson and gamma*
$$\pi(\theta|x) = b^a \theta^{a-1} e^{-b\theta}/\Gamma(a) \quad (\theta > 0). \tag{10}$$

D 4. *Normal.* Here $\theta = (\mu, \sigma)$ but to make the integration problems that arise more directly manageable it is convenient to introduce $\tau = 1/\sigma$ and thus to take $\theta = (\mu, \tau)$. We then suppose that $\pi(\theta|x)$ is of normal–gamma type
$$\pi(\theta|x) \propto \tau \exp\{-\tfrac{1}{2}b\tau^2(\mu-a)^2\}\tau^{w-1}\exp(-\tfrac{1}{2}wv\tau^2). \tag{11}$$
It is worth observing here that this does in fact cover the regression experiment case (*b*). The fact that, for fixed z, the parameters α and β appear only in the form $\mu = \alpha + \beta z$, and that a prior multivariate normal–gamma density (see Raiffa & Schlaifer, 1961, chapter 3 for definition and suitability) on (α, β) and τ induces a posterior normal–gamma density, which in turn 'condenses' into a posterior normal–gamma density, say $\pi(\mu, \tau|x)$ for μ and τ, allows the treatment within the framework of D 4.

From these it is a matter of routine summation or integration as in (1) to obtain the following densities for $\pi(y|x)$.

D 1. *Binomial*
$$\pi(y|x) = \binom{l}{y} \frac{B(a+y, b+l-y)}{B(a,b)}. \tag{12}$$

D 2. *Poisson*
$$\pi(y|x) = \frac{\Gamma(a+y)}{y!\,\Gamma(a)} \left(\frac{l}{b+l}\right)^y \left(\frac{b}{b+l}\right)^a. \tag{13}$$

D 3. *Gamma*
$$\pi(y|x) = \frac{b^a}{B(a,l)} \frac{y^{l-1}}{(b+y)^{a+l}}. \tag{14}$$

D 4. *Normal*
$$\pi(y|x) = \frac{w^{\frac{1}{2}w}\{b/v(1+b)\}^{\frac{1}{2}}}{B(\frac{1}{2}, \frac{1}{2}w)\{w + b(y-a)^2/v(1+b)\}^{\frac{1}{2}(w+1)}}. \tag{15}$$

4. Standard inference predictions

The standard results on prediction in text-books and journals are invariably of inference type and fall into two categories, one involving a simple V specification and the other a simple U specification. These may be termed *expected cover* predictions and *tolerance region* predictions, respectively. It is the simplicity of the V and U functions which allows (7) and (8) to be transformed into probabilistic statements which are the usual starting points of such analysis.

Expected cover predictions

These predictions use

$$V^*(R,y) = \begin{cases} 1 & (y \in R), \\ 0 & (y \notin R), \end{cases} \tag{16}$$

which, through (4) gives $U(R,\theta) = P_F(R|\theta)$, the *cover* of R at θ. Relation (7) can be written in the form

$$P_{EF}\{(x,y): y \in \delta(x)|\theta\} = q \tag{17}$$

for all θ. A frequentist difficulty of a slightly subtler form than those discussed in Aitchison (1964) is now apparent. For in the usual frequency interpretation of probability (17) is meaningful in terms of repetitions of E *and* F, an inference being made after each E, and its success or failure being assessed against the F following that E, so that F 1 is the appropriate condition. We saw, however, in § 1 that for the frequentist who wished to argue from a V specification it was necessary to think in terms of case 2 where E served for a series of future F's, for which F 2 is the appropriate frequentist condition. It would therefore seem that the frequentist cannot happily consider a V-specification except, of course, by some appeal to F 3. Nevertheless, this does leave open the possibility of starting the argument at the corresponding U or asserting that (17) is the basic inferential statement, within the terms of F 1 or F 3.

Such frequentist expected cover regions have been considered usually in terms of relationship (17) by various writers; see for example, Thatcher (1964) who treats the binomial case, Proschan (1953) and Fraser & Guttman (1956), who deal with the normal case. The interval associated with the normal regression case is commonly quoted in text-books as a prediction interval; see, for example, Bowker & Liebermann (1959). The technical problem of extending this type of normal regression inference prediction to the case where the future experiment is to be performed at a finite known set of z values is resolved by Lieberman (1961).

One technical point worth making here is that $\delta(x)$ satisfying (17) is in no sense a frequentist *confidence* region for y; the statement (17) involves the joint distribution of (x,y) and so y is not being considered as a 'parameter'.

Bayesian analysis requires the choice of an R such that

$$\int_R \pi(y|x)\,dy = q \tag{18}$$

and so R is essentially a Bayesian confidence region. Thatcher (1964) deals with the binomial case and Lindley (1965, pp. 212–13) uses this as the basis for his prediction interval for the case of normal linear regression.

Since we have had difficulty in tracing any treatment of the Poisson and gamma cases we have thought it convenient to present in Table 1 the Bayesian and frequentist intervals for the four cases. We have not included derivations of the frequentist results which for D 1–D 3 can be obtained along the lines of Thatcher (1964) and involve routine but tedious summation and integration. The Bayesian results are readily obtained in terms of tabulated functions from (18) and the formulae (12)–(15).

Tolerance region predictions

The specification here is

$$U^*(R,\theta) = \begin{cases} 1 & \text{if} \quad P_F(R|\theta) \geqslant c, \\ 0 & \text{if} \quad P_F(R|\theta) < c, \end{cases} \tag{19}$$

where c is a specified desired cover, and leads either to Bayesian or frequentist tolerance regions. Since these have been discussed at length by Aitchison (1964) we confine comment here to observing that the formulation suffers from the serious practical difficulty of deciding on a balance between c and q. In the form of words commonly used in such problems, is it better to be 99 % certain ($q = 0{\cdot}99$) that a predictor will provide 95 % cover or to be 95 % certain of 99 % cover?

Although Aitchison (1964) does not discuss tolerance regions in the regression situation there is no technical difficulty in carrying out this extension through the device discussed in density specification D 4. Wallis (1951) has already considered the frequentist aspect of this extension.

Table 1. *Bayesian and frequentist expected cover predictions*

Density	Type of interval (end values included)	Bayesian	Frequentist
D 1	$[0, \delta(x)]$	$\delta(x) = \min\{\gamma\colon P(a+b+l-1, \gamma, a+b-1, a-1) \leqslant 1-q\} - a$	$a = x+1, b = k-x$
	$[\delta(x), l]$	$\delta(x) = \max\{\gamma\colon P(a+b+l-1, \gamma, a+b-1, a-1) \geqslant q\} - a + 1$	$a = x, b = k-x+1$
D 2	$[0, \delta(x)]$	$\delta(x) = \min\{\gamma\colon I_{b/(b+1)}(a, \gamma) \geqslant q\} - 1$	$a = x+1, b = k$
	$[\delta(x), \infty)$	$\delta(x) = \max\{\gamma\colon I_{b/(b+1)}(a, \gamma) \leqslant 1-q\}$	$a = x, b = k$
D 3	$[0, \delta(x)]$	$\delta(x) = b\{1 - B(a, l; 1-q)\}/B(a, l; 1-q)$	$a = k, b = x$
	$[\delta(x), \infty)$	$\delta(x) = b\{1 - B(a, l; q)\}/B(a, l; q)$	$a = k, b = x$
D 4	$[0, \delta(x)]$	$\delta(x) = a + \{v(1+b)/b\}^{\frac{1}{2}} t(w; q)$	$a = m$
	$[\delta(x), \infty)$	$\delta(x) = a - \{v(1+b)/b\}^{\frac{1}{2}} t(w; q)$	$b = 1/h$
	$[\delta_1(x), \delta_2(x)]$ $\begin{Bmatrix}\delta_1(x)\\ \delta_2(x)\end{Bmatrix}$	$= a \mp \{v(1+b)/b\}^{\frac{1}{2}} t\{w; \frac{1}{2}(q+1)\}$	$\begin{cases} v = s \\ w = \nu \end{cases}$

The notation $\min\{\gamma\colon A\}$ is used to denote the least γ satisfying property A.

The function P is the hypergeometric distribution function as tabulated by Lieberman & Owen (1961).

The function I is the incomplete beta function as tabulated by Pearson (1934).

$B(a, l; q)$ is the q-fractile of the incomplete beta distribution, and satisfies $I_{B(a,l;q)}(a, l) = q$. It is thus obtainable from Pearson (1934).

$t(w; q)$ is the q-fractile of the $t(w)$ distribution.

5. Bayesian decision predictions

As we have already observed, the specifications V^* and U^* of the preceding section are too naïve for decision problems. While such simple specifications are not directly useful it is, however, possible to build on them to obtain procedures for decision predictions on the basis of value specifications. Our interest will centre on specifications which are of the form

$$V_j(r, y) = \begin{cases} y^{(j)} = y(y-1)\ldots(y-j+1) & (y \leqslant r), \\ 0 & (y > r), \end{cases} \tag{20}$$

for density specifications D 1 and D 2;

$$V_j(r, y) = \begin{cases} y^j & (y \leqslant r), \\ 0 & (y > r), \end{cases} \tag{21}$$

for D 3; and

$$V_j(r, y) = \begin{cases} (y-a)^j & (y \leqslant r), \\ 0 & (y > r), \end{cases} \tag{22}$$

for D 4. From these it is possible to construct more realistic value functions, for example, in the demand and supply problem outlined in § 1 the prediction region R is of the form $(-\infty, r]$ or $[0, r]$; if there is a cost $C(r)$ of operating R, which should usually include an

allowance for depreciation, and if the penalties for supply exceeding, or falling short of demand are proportional to the quantity in excess, or unsupplied, then a sensible value specification $V(R,y)$ or $V(r,y)$ is

$$V(r,y) = \begin{cases} \alpha(y-r)-C(r) & (y \leqslant r), \\ \beta(r-y)-C(r) & (y>r). \end{cases} \tag{23}$$

Then V may be easily expressed in terms of the simpler V_j specifications as

$$V(r,y) = \alpha V_1(r,y) - \alpha r V_0(r,y) + \beta r\{1-V_0(r,y)\} - \beta\{V_1(\infty,y)-V_1(r,y)\} - C(r) \tag{24}$$

for D 1–D 3, and as

$$V(r,y) = \alpha V_1(r,y) - \alpha(r-a)V_0(r,y) + \beta(r-a)\{1-V_0(r,y)\} - \beta\{V_1(\infty,y)-V_1(r,y)\} - C(r) \tag{25}$$

for D 4.

Again, when a finite interval $R = (r_1, r_2)$ is required and when losses are quadratic in distance of y from the interval, and increase quadratically with distance inside the interval (the deeper y is contained in the interval the more we may have used too large a prediction interval) we may have a V-specification

$$V(R,y) = \begin{cases} -(y-r_1)^2 & (y \leqslant r_1), \\ -(y-r_1)(r_2-y) & (r_1 < y < r_2), \\ -(y-r_2)^2 & (y \geqslant r_2). \end{cases} \tag{26}$$

This is again easily expressed in terms of V_j functions; for example, for problem D 3,

$$\begin{aligned} V(R,y) = &-2V_2(r_1) + (3r_1+r_2)V_1(r_1) - r_1(r_1+r_2)V_0(r_1) \\ &+ 2V_2(r_2) - (r_1+3r_2)V_1(r_2) + r_2(r_1+r_2)V_0(r_2) \\ &+ 2r_2 V_1(\infty) - r_2^2 - V_2(\infty) \end{aligned} \tag{27}$$

in an obvious abbreviated notation.

More complicated polynomial value specifications can obviously also be expressed in terms of these simpler V_j's.

Moreover, integrating with respect to $\pi(y|x)$ to obtain $H(r,x)$ or briefly $H(r)$ as in (2) is a linear operation and so the V's in the above formulae may be replaced by the corresponding H's. It is therefore useful and sensible to have a catalogue of $H_j(r)$ corresponding to the simple V_j value functions for the four standard density specifications. These are given in Table 2; see footnote to Table 1 for definitions of the B, I and P notation.

We note in passing that, when a 'point' prediction, i.e. a single predicted outcome r of F, is required, the value specifications $V(r,y) = -|y-r|$ and $V(r,y) = -(y-r)^2$ lead to the use of the median and the mean of the $\pi(y|x)$ density.

One general comment is worth making here. The $H_j(r)$ are easily seen to be the jth order incomplete moments, over $(-\infty, r)$ or $(0, r)$ instead of the complete space Y, associated with the $\pi(y|x)$ distribution; these are factorial moments for D 1 and D 2, moments about the origin for D 3 and moments about a for D 4. Such incomplete moments occur in allied decision problems, for example, those of Raiffa & Schlaifer (1961, chapter 6). The Raiffa & Schlaifer problem is concerned with situations where the state or parameter space coincides with the decision space (as in estimation problems) or can be made by transformation to coincide with it. In our type of problem the decision space is Y or $\mathscr{Y}$ and the state space is

$Y \times \Theta$, θ being a nuisance parameter in the sense that the value specification does not depend on it. While there is thus a certain similarity between the two types of problem the dependence of the present situation on θ is too strong to reduce it to the Raiffa & Schlaifer problem. This dependence on θ arises from the fact that the necessary relationship between the informative experiment E and the future experiment F can only satisfactorily be described in terms of their common indexing parameter θ. In order to apply the Raiffa & Schlaifer theory to our problems we would have to be able to express the prior uncertainty about the outcome of F directly as a density $\pi(y)$ on Y and to allow the information x to alter this directly to $\pi(y|x)$ without any reference to θ. To do this we would really have to think of the family of densities describing E as being of the form $p(x|y)$, i.e. indexed by $y \in Y$, and this is seldom a natural view. The necessary introduction of θ leads to a more involved type of decision analysis but the resulting distribution theory is no more complicated.

Table 2. $H_j(r)$ *for the densities* D 1–D 4

Density	$H_j(r)$
D 1	$\frac{l^{(j)}B(a+j,b)}{B(a,b)}P(a+b+l-1,\ a+r,\ l-j,\ r-j)$
D 2	$\left(\frac{l}{b}\right)^j \frac{\Gamma(a+j)}{\Gamma(a)} I_{b/(b+l)}(a+j, r-j+1)$
D 3	$\frac{b^j B(l+j, a-j)}{B(l,a)} I_{r/(b+r)}(l+j, a-j)$
D 4	$\frac{1}{2}\left\{\frac{vw(1+b)}{b}\right\}^{\frac{1}{2}j} \frac{B\{\frac{1}{2}(j+1), \frac{1}{2}(w-j)\}}{B(\frac{1}{2}, \frac{1}{2}w)} [(-1)^j + I_g\{\frac{1}{2}(j+1), \frac{1}{2}(w-j)\}] \quad (r>a),$
	$(-1)^j \frac{1}{2}\left\{\frac{vw(1+b)}{b}\right\}^{\frac{1}{2}j} \frac{B\{\frac{1}{2}(j+1), \frac{1}{2}(w-j)\}}{B(\frac{1}{2}, \frac{1}{2}w)} [(-1)^j - I_g\{\frac{1}{2}(j+1), \frac{1}{2}(w-j)\}] \quad (r \leqslant a)$

where $$g = \frac{(r-a)^2}{\{vw(1+b)/b\} + (r-a)^2}.$$

Table 2 provides a convenient means of constructing $H(R, x)$ for a particular problem. The method of subsequently maximizing $H(\cdot, x)$ must depend largely on its form and on the computing facilities available. Therefore instead of entering a general, and necessarily vague, discussion on the relative merits of differentiation, iterative and search techniques, we provide an illustrative example which demonstrates a technique that we have found useful in hospital engineering supply problems. We there consider the value specification (23). From (24) we have, for D 3 and omitting the x in the notation for brevity,

$$H(r) = (\alpha+\beta) H_1(r) - (\alpha+\beta) r H_0(r) + \beta r - \beta H_1(\infty) - C(r). \tag{28}$$

Setting $H'(r) = 0$ and using the fact that $H_j'(r) = r^j \pi(r|x)$ we have as condition for a maximum

$$(\alpha+\beta) H_0(r) + C'(r) - \beta = 0 \tag{29}$$

and standard computational techniques may be applied to obtain a solution.

The specification of α, β and $C(r)$ is necessary for a completely satisfactory decision analysis. When the cost $C(r)$ of operating the system is small (for example, when the cost is mainly capital outlay and the system is to operate over a long period) or if the engineer judges the cost to be of secondary importance compared with the failure–success assessment, then we may set $C(r) = 0$ and obtain the following simple analysis. It is now easier to work in terms of the *relative cost ratio* $\lambda = \beta/\alpha$, which characterizes the relative costs of failure

and success, and to take

$$V_\lambda(r,y) = \begin{cases} y-r & (y \leqslant r), \\ \lambda(r-y) & (y>r). \end{cases} \tag{30}$$

For the engineer who hesitates to specify λ it is possible to display the dependence on λ of the *design value* r_λ which maximizes $H_\lambda(\cdot)$ by plotting the graph (λ, r_λ). From (29) the relation between r_λ and λ is

$$(1+\lambda)\,H_0(r_\lambda)-\lambda = 0 \tag{31}$$

and the graph is very easily plotted by varying r_λ and obtaining the corresponding λ, given by

$$\lambda = H_0(r_\lambda)/\{1-H_0(r_\lambda)\}. \tag{32}$$

The corresponding expected utility is given by

$$\begin{aligned} H(r_\lambda) &= (1+\lambda)\,H_1(r_\lambda)-\lambda H_1(\infty)-r\{(1+\lambda)\,H_0(r_\lambda)-\lambda\} \\ &= \frac{H_1(r_\lambda)-H_1(\infty)\,H_0(r_\lambda)}{1-H_0(r_\lambda)} \end{aligned} \tag{33}$$

and the graph $\{\lambda, H(r_\lambda)\}$ can be plotted at the same time. To illustrate this point Fig. 2 shows the two graphs for case D 3 when F is $\mathrm{Ga}(1,\theta)$ and the outcome of E leads to $a = 25$, $b = 50$. The feature of this graph, that it shows immediately the increase in λ which is catered for by a specified increase in the supply quantity r and the corresponding change in expected utility, seems to be particularly appealing to some engineers. This is especially so when the possible values of r increase by discrete steps, for example, in case of determining cable and pipe sizes; then it is almost more meaningful to quote the range of λ values for which a particular r is optimal.

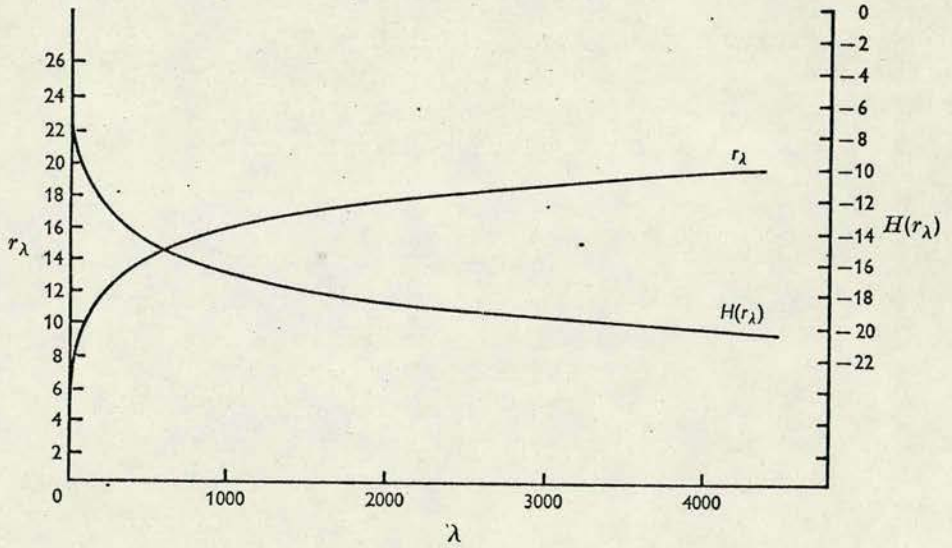

Fig. 2. Graphs of r_λ, the design value corresponding to a relative cost ratio λ, and $H(r_\lambda)$, the maximum expected utility obtained by use of r_λ.

6. PREDICTIONS FOR CLASSES OF EXPERIMENTS

There are a number of prediction problems in which it is convenient to consider a whole class $\mathscr{F}$ of possible future experiments F_z, indexed by $z \in Z$, i.e.

$$\mathscr{F} = \{F_z : z \in Z\}.$$

A typical situation is where the informative experiment E is a regression experiment

$(F_{z_1}, \ldots, F_{z_n})$, and it is not known precisely for how many, or which, future experiments predictions will be required. It is assumed, however, at least in the subsequent three subsections that the appropriate z is known at the time of making any required prediction. Since it is now essential to show the dependence of the experiment F_z on z we use $p_F(y|z,\theta)$ for the density associated with F_z, and $P_F(\cdot|z,\theta)$ for the associated probability measure. Several analyses of this situation are possible; the one chosen should, of course, reflect the realities of the situation as closely as possible.

(1) *Simultaneous tolerance regions*

Liebermann & Miller (1963) have suggested one resolution of the problem through a concept of simultaneous tolerance regions by containing all the possible predictions within one probabilistic statement. The approach is a frequentist one and since a predictor

$$\delta_z: x \to \delta_z(x) \in \mathscr{Y}$$

is required for each possible z, what is required in order to define a whole prediction precedure is a whole class of predictors

$$\Delta = \{\delta_z: z \in Z\}.$$

The Liebermann & Miller procedure, which is for the normal regression situation, is based on a modification of the simple tolerance region probabilistic statement (see, for example, Aitchison, 1964) to

$$P_E[x: P_F\{\delta_z(x)|z,\theta\} \geqslant c \quad \text{for all} \quad z \in Z|\theta] = q \quad \text{for all} \quad \theta \in \Theta. \tag{34}$$

(Actually, there is a further modification in their work to allow the possibility of c varying over some class, but the simpler version (34) is sufficient illustration here. The statement is clearly designed for a situation where it is envisaged that repeated use will be made of the one outcome of the informative experiment to provide answers to all sorts of prediction questions about different experiments in the class $\mathscr{F}$. The sense of added responsibility arising from this extended use of the information instils a feeling of extreme caution in the statistician and he seeks to play safe by constraining his entire procedure within the probability value q. It is also clear (see the argument in the next subsection) that the containment within the probabilistic statement (34) is motivated by the feeling that the (unspecified) utilities are not additive over inference or decisions. If such is the case then (34) can be regarded as a useful technique for those who are faced with a series of similar prediction problems (as in F 2) based on repetitions of E, in which case q has a relative frequency interpretation. Otherwise, an appeal to F 3 is necessary to make the process acceptable. It should be pointed out here that the difficulty of counterbalancing q and c in this much more involved situation seems to require tremendous insight.

(2) *Frequentist predictions when utilities are additive over predictions*

We assume that when a number of predictions are to be made on the same information the utilities are additive over predictions. The following argument can then be applied in the frequentist analysis. There is some $\tilde{\omega}(z)$, albeit completely unknown, which describes the relative frequency with which F_z actually arises. Since $U\{\delta_z(x), z, \theta\}$ measures the utility of Δ in predicting for F_z when E yields x, it is reasonable, because of the additive utility property, to assess the overall merits of Δ in relation to $\mathscr{F}$ by the average utility per forecast, viz.

$$I(\Delta, x, \theta) = \int_Z dz\, \tilde{\omega}(z)\, U\{\delta_z(x), z, \theta\}.$$

Using this I as a utility function in a straightforward frequentist inference or decision

analysis we are led to seek a procedure Δ which maximizes or sets equal to q the average

$$I(\Delta, \theta) = \int_X dx\, p_E(x|\theta) \int_Z dz\, \tilde{\omega}(z)\, U\{\delta_z(x), z, \theta\} \tag{35}$$

for all $\theta \in \Theta$. Reordering this double integration, a step which will in general be possible, we obtain

$$\begin{aligned} I(\Delta, \theta) &= \int_Z dz\, \tilde{\omega}(z) \int_X dx\, p_E(x|\theta)\, U\{\delta_z(x), z, \theta\} \\ &= \int_Z dz\, \tilde{\omega}(z)\, G\{\delta_z, z, \theta\} \end{aligned} \tag{36}$$

and so, to obtain a Δ which, for instance, maximizes $I(\Delta, \theta)$, all that is necessary is to find *for each fixed* z, a predictor δ_z which maximizes $G(\delta_z, z, \theta)$. In other words, we are faced with the case of a specified future experiment as in § 2. No principle of simultaneity in this situation seems necessary even on this frequentist view of the problem. Lack of knowledge of $\tilde{\omega}(z)$ has been placed in much the same category as not knowing the true parameter value in more straightforward frequentist analyses.

(3) *Bayesian predictions for classes of experiments*

When the Bayesian knows the outcome x from E and the F_z for which he has to predict then his information about y is contained in the density

$$\pi(y|x, z) = \int_\Theta p_F(y|z, \theta)\, \pi(\theta|x)\, d\theta. \tag{37}$$

He will then base his prediction region R_z on

$$H(R_z, x, z) = \int_Y V(R_z, y)\, \pi(y|x, z)\, dy, \tag{38}$$

either maximizing with respect to R_z in decision prediction or setting $H(R_z, x, z) = q$ for inference prediction. It is thus clear that there is no difference here from the ordinary regression problem with fixed F_z, and so we need make no further comments on the problem.

(4) *Predictions when z is not known at the time of prediction*

We consider only the case where utilities are additive over predictions. When the particular F_z is not known it is required to provide a single region R (not depending on z) to be used for any F_z which may arise. For example, in a demand and supply problem the demand y for a commodity at an operation may depend on a variable temperature z and the supply system has to be determined in advance; e.g. it may involve laying of pipes or cables. We suppose that we have information about the demands at various temperature levels in the past, that is the outcomes $y_1, \ldots, y_n$ from experiments $F_{z_1}, \ldots, F_{z_n}$. Clearly we require to have information about the likely temperatures that may arise before we can effectively predict R. The approach we adopt is essentially Bayesian, and proceeds as follows. If the information about z contained in $\tilde{\omega}(z)$ is independent of the Bayesian information about θ, and so about y, contained in $\pi(\theta|x)$, or $\pi(y|x, z)$—we shall presently explore more deeply how this independence may arise—then for decision predictions the Bayesian will attempt to maximize

$$J(R, x) = \int_Z \tilde{\omega}(z)\, H(R, x, z)\, dz. \tag{39}$$

It may reasonably be asked how we can assume that the information available on z and θ after the informative experiment can be regarded as independent—with density $\tilde{\omega}(z)\, \pi(\theta|x)$ —when, in the regression situation indicated earlier, the information about the relative

frequencies of future z's is also to be found in $x = (z_1, y_1; \ldots; z_n, y_n)$. The density associated with the joint occurrence of z and y can be expressed as

$$p(z|\tau)\, p_E(y|z, \theta),$$

where $p(z|\tau)$ is the marginal density of z and depends on a parameter τ. Since this separation is often the natural way of regarding the informative experiment it may well be that the prior information about τ and θ is independent in the sense that the joint prior density is $\pi_1(\tau)\,\pi_2(\theta)$. If this is the case then a trivial application of Bayes's theorem to obtain the posterior density of τ and θ shows that the factorization is retained; the posterior density is of the form

$$\pi_1(\tau|x)\,\pi_2(\theta|x).$$

It is the variation in future z which is relevant to our determination of R and so we must convert this to posterior information about z and θ with density

$$\int d\tau\, p(z|\tau)\,\pi_1(\tau|x)\,\pi_2(\theta|x) = \tilde{\omega}(z|x)\,\pi_2(\theta|x);$$

again the factorization persists and the assumption originally made about independence is thus valid in these circumstances.

If, of course, the experimenter has himself selected the z values of the informative experiment E then the Bayesian assessment of $\tilde{\omega}(z)$ must be made on information outside E and so again the assumption of independence seems reasonable.

The integral (or sum in the case of discrete Z) $J(R, x)$ is not expressible in simple terms, but with the facility for constructing $H(R, x, z)$ described in § 5, the problem of maximizing $J(\cdot, x)$ is within the scope of most automatic computers.

We wish to thank the referees for their very constructive comments on an earlier draft of this paper. We also wish to express our thanks to members of the Hospital Engineering Research Unit of the University of Glasgow for presenting us with problems which stimulated this research.

References

Aitchison, J. (1964). Bayesian tolerance regions. *J.R. Statist. Soc.* B, **26**, 161–75; with discussion, 192–210.

Bowker, A. H. & Liebermann, G. J. (1959). *Engineering Statistics*. Englewood Cliffs: Prentice-Hall.

Fraser, D. A. S. & Guttman, I. (1956). Tolerance regions. *Ann. Math. Statist.* **27**, 162–71.

Liebermann, G. J. (1961). Prediction regions for several predictions from a single regression analysis. *Technometrics*, **3**, 21–7.

Liebermann, G. J. & Miller, R. G. (1963). Simultaneous tolerance intervals in regression. *Biometrika*, **50**, 155–68.

Liebermann, G. J. & Owen, D. B. (1961). *Tables of the Hypergeometric Probability Distribution*. Stanford University Press.

Lindley, D. V. (1965). *Introduction to Probability and Statistics*. Part 2. *Inference*. Cambridge University Press.

Neyman, J. (1964). Contribution to discussion on papers on fiducial probability. *Bull. Int. Stat. Inst.* **40**, 927–32.

Pearson, K. (1934). *Tables of the Incomplete Beta-Function*. Cambridge University Press.

Proschan, F. (1953). Confidence and tolerance intervals for the normal distribution. *J. Amer. Statist. Ass.* **44**, 550–64.

Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University.

Thatcher, R. (1964). Relationships between Bayesian and confidence limits for predictions. *J.R. Statist. Soc.* B, **26**, 176–92; with discussion 192–210.

Wallis, W. A. (1951). Tolerance intervals for linear regression. *Proc. Second. Berkeley Symp.* Berkeley: University of California Press.

AITCHISON, J. (1966)

Expected-cover and linear-utility tolerance intervals

Reprinted from *J. R. Statist. Soc.* B28, 57-62

Expected-cover and Linear-utility Tolerance Intervals

By J. AITCHISON

University of Liverpool

[Received May 1965. Revised November 1965]

SUMMARY

Expected-cover tolerance intervals are commonly used as a means of statistical prediction. In their construction the effectiveness of such intervals in relation to a future observation y is assessed only in terms of "success" if y falls in the interval and of "failure" otherwise; no account is taken in the assessment of how far inside or outside the interval y happens to fall. The present paper considers the construction of intervals which do take account of this factor in a linear way. From a Bayesian viewpoint it is found that expected-cover and linear-utility intervals can be regarded as equivalent through a simple relation between the expected cover and the relative cost ratio. For the frequentist approach it is first shown that linear-utility intervals can be simply constructed for the normal and gamma distributions. Comparison of these with expected-cover intervals shows that, while there is not complete identity, there is an equivalence in a "large-sample" sense.

1. INTRODUCTION

TOLERANCE intervals are a means of statistical prediction. From a set x of observations —the outcome of some informative experiment E—a tolerance interval $\delta(x)$—a subset of the outcome space of some future experiment F—is constructed. The choice of the predictor δ or the prediction $\delta(x)$ is in some probabilistic sense aimed at containing the outcome y of F or a sequence of such outcomes of repetitions of F, and the precise sense in which $\delta(x)$ is related to F determines the type of tolerance interval. In this paper we study, from both a frequentist and a Bayesian approach, two such types—expected-cover and linear-utility intervals—and their interrelations.

Suppose that F is described by a density $p_F(y \mid \theta)$ $(y \in Y)$, where the true value of the indexing parameter θ is not known; we denote by $P_F(\cdot \mid \theta)$ the corresponding probability measure. Suppose further that an *independent* (for given θ) informative experiment E is described by a density $p_E(x \mid \theta)$ $(x \in X)$, the common parameter θ ensuring the possibility of prediction; again, $P_E(\cdot \mid \theta)$ is the corresponding probability measure. We denote the product probability measure which describes E and F by $P_{EF}(\cdot \mid \theta)$. A frequentist expected-cover tolerance interval $\delta(x)$ depends on the choice of a predictor δ satisfying the probabilistic statement

$$P_{EF}\{(x, y): y \in \delta(x) \mid \theta\} = q \quad \text{for all } \theta, \tag{1}$$

where q is the "confidence coefficient"; see, for example, Proschan (1953), Fraser and Guttman (1956), Bowker and Lieberman (1959, pp. 253–257), Thatcher (1964). This probabilistic statement is equivalent to the relation

$$\int_X P_F\{\delta(x) \mid \theta\} p_E(x \mid \theta)\, dx = q \quad \text{for all } \theta. \tag{2}$$

Since $P_F\{\delta(x)|\,\theta\}$ is the *cover* provided by $\delta(x)$ when the true parameter value is θ, the statement (2) shows the origin of the name "expected-cover".

For a Bayesian expected-cover interval we require to find a prediction $R = \delta(x)$ satisfying

$$\Pi(R|\,x) = q, \tag{3}$$

where $\Pi(\cdot\,|\,x)$ is the probability measure corresponding to the posterior density $\pi(y|\,x)$ of y for given x; see Aitchison and Sculthorpe (1965) for a discussion of the appropriateness of this density.

Aitchison (1964, particularly in the discussion) and Aitchison and Sculthorpe (1965) have set such problems within the framework of a general inference and decision approach to prediction. If $V(R, y)$ denotes the value (or y-utility) of an interval or region R when y is the observed outcome of F, and if

$$U(R, \theta) = \int_Y V(R, y) p_F(y|\,\theta)\, dy, \tag{4}$$

then the frequentist decision approach leads to the search for a predictor δ which maximizes

$$G(\delta, \theta) = \int_X U\{\delta(x), \theta\} p_E(x|\,\theta)\, dx \tag{5}$$

for all θ. The corresponding Bayesian decision prediction approach requires the determining of an R which maximizes

$$H(R, x) = \int_Y V(R, y)\, \pi(y|\,x)\, dy. \tag{6}$$

For an expected-cover analysis we take

$$V(R, y) = \begin{cases} 1 & (y \in R), \\ 0 & (y \notin R), \end{cases} \tag{7}$$

from which, by (4), $U(R, \theta) = P_F(R|\,\theta)$, so that $G(\delta, \theta)$ is the left-hand side of (2), and so of (1). Similarly, $H(R, x)$ becomes the left-hand side of (3). That we do not attempt to maximize here is a consequence of the fact that $\delta(x) = Y$ gives the maximum of 1 for G and for H, and this leads to the trivial prediction that $y \in Y$. Instead of maximizing we transform the problem from one of decision type to essentially that of inference by seeking a δ such that

$$G(\delta, \theta) = q \quad \text{for all } q, \tag{8}$$

or an R such that

$$H(R, x) = q, \tag{9}$$

where $q < 1$ is a preassigned number.

The V-function which is the basis of expected-cover analysis is so simple that it leads to inference rather than a true decision problem requiring maximization. This simplicity arises from the fact that the expected-cover V-assessment treats as equally serious any y outside R and as equally successful any y inside R. While the V-specification may in practice be a crude assessment of the worth of an interval or

predictor, the question immediately arises: is there some more sophisticated assessment which takes account of the extent by which y falls outside or inside R and leads to equally simple constructions? One such assessment for an interval of type $R = (-\infty, r)$ is the linear y-utility

$$V(r,y) = \begin{cases} y-r & (y \leqslant r), \\ \lambda(r-y) & (y > r). \end{cases} \tag{10}$$

Here the loss involved if y falls outside the interval is considered as proportional to the distance $y-r$ of y from the interval. Moreover, if y falls inside the interval we have used a larger interval (larger by an amount $r-y$) than was necessary and the loss specified by V is proportional to this amount. The choice of the factors of proportionality as λ and 1 loses no generality, and allows the two types of loss to be treated as of unequal seriousness. Note that the *relative cost ratio* λ is not necessarily greater than 1. Aitchison and Sculthorpe (1965) have discussed the suitability of such linear value functions to the design of certain supply systems, where not only must there be a penalty for undersupply but also a penalty for overdesign. An alternative view of such problems is, of course, that they are aimed at estimating the "parameter" y. It is from the estimation aspect that Raiffa and Schlaifer (1961, Chapter 6) use the same linear utility. Their view of the problem is, however, essentially different in that they feel able to describe the situation in terms only of the "parameter" y without reference to the natural indexing parameter θ.

An interesting and immediate consequence of relations (18) and (32) of Aitchison and Sculthorpe (1965), not made explicit by them, is that Bayesian expected-cover and linear-utility intervals, of the form $(-\infty, r)$ or $(0, r)$, are exactly the same provided

$$q = \lambda/(1+\lambda). \tag{11}$$

The provision of this alternative view of expected-cover intervals may make them more attractive to some users in that they feel that it is easier to assess the relative cost factor λ than the more nebulous "confidence coefficient" q.

We now explore how far it is possible to develop the frequentist linear-utility approach; since in general there is no guarantee even of the existence of a function δ satisfying (8) we first investigate the often amenable normal case.

2. Frequentist Linear-utility Normal Tolerance Intervals

In the normal case we take $p_F(y \mid \theta)$ to be a $N(\mu, \sigma^2)$ density with mean μ and standard deviation σ—so that $\theta = (\mu, \sigma^2)$. We suppose that the informative experiment E provides us with independent estimating statistics m and s for μ and σ—so that $x = (m, s)$—and that m is $N(\mu, h\sigma^2)$ and $\nu s^2/\sigma^2$ is $\chi^2(\nu)$, that is, χ^2 with ν degrees of freedom. This description includes the case where E is n replicates of F, and also the case where E is a normal regression experiment, carried out say at levels $z_1, \ldots, z_n$ of a predetermined variable, F being a future experiment to be carried out at z; see, for example, Wallis (1951), Owen (1963), Aitchison and Sculthorpe (1965). In such circumstances it is well known that a tolerance interval of expected cover q is $(-\infty, m+k_0 s)$, with

$$k_0 = \sqrt{(1+h)}\, t(\nu; q), \tag{12}$$

where $t(\nu; q)$ denotes the q-fractile of the t distribution with ν degrees of freedom.

It will pay us to retain the general notation of Section 1 while investigating the U-function corresponding to (10). From (4) we have

$$U(r,\theta)=\int_{-\infty}^{r}(y-r)P_F(y|\theta)\,dy+\lambda\int_{r}^{\infty}(r-y)P_F(y|\theta)\,dy.$$

Later in our analysis we shall require

$$\frac{\partial U(r,\theta)}{\partial r}=\lambda-(1+\lambda)P_F(r|\theta), \tag{13}$$

where we now use $P_F(r|\theta)$ as an abbreviation for $P_F\{(-\infty,r)|\theta\}$.

Just as in expected-cover analysis it seems sensible to consider limits of the form $r=\delta(m,s)=m+ks$, based on (m,s) which are joint sufficient statistics for (μ,σ), so we consider here limits of this type, and attempt to find a k, if any, which maximizes $G(k,\theta)$ for all θ; note that we write $G(k,\theta)$ for $G(\delta,\theta)$ since k completely specifies δ. The joint density of (m,s) is

$$p(m,s)=\frac{1}{\sigma\sqrt{(2\pi h)}}\exp\left(-\frac{(m-\mu)^2}{2h\sigma^2}\right)\frac{\nu^{\frac{1}{2}\nu}}{2^{\frac{1}{2}\nu-1}\Gamma(\frac{1}{2}\nu)}\frac{s^{\nu-1}}{\sigma^{\nu}}\exp\left(-\frac{\nu s^2}{2\sigma^2}\right),$$

and (5) gives

$$G(k,\theta)=\int_{-\infty}^{\infty}\int_{0}^{\infty}U(m+ks,\theta)\,p(m,s)\,dm\,ds. \tag{14}$$

It is fairly easy to establish that a maximizing value of k necessarily occurs where the derivative of $G(k,\theta)$ with respect to k is zero. From (13) the derivative equation is obtained as

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\{\lambda-(1+\lambda)P_F(m+ks|\theta)\}\,sp(m,s)\,dm\,ds=0. \tag{15}$$

Now

$$P_F(m+ks|\theta)=\Phi\left(\frac{m-\mu}{\sigma}+k\frac{s}{\sigma}\right),$$

where Φ is the distribution function of the standardized normal distribution. If we introduce the change of variables

$$M=\frac{m-\mu}{\sigma\sqrt{h}},\quad S=\frac{s\sqrt{\nu}}{\sigma\sqrt{(\nu+1)}}, \tag{16}$$

we reduce the left-hand side of (15), after the cancellation of a factor, to

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}[\lambda-(1+\lambda)\Phi\{M\sqrt{h}+kS\sqrt{(\nu+1)}/\sqrt{\nu}\}]f(M)\,g(S)\,dM\,dS, \tag{17}$$

where $f(M)$ and $g(S)$ are $N(0,1)$ and $\{\chi^2(\nu+1)/(\nu+1)\}^{\frac{1}{2}}$ densities respectively. In considering (17) we can therefore treat M and S as independent with the specified densities and if we introduce another $N(0,1)$ variable W, say, independent of (M,S), we see that the derivative equation yields

$$\Pr\{W\leqslant M\sqrt{h}+kS\sqrt{(\nu+1)}/\sqrt{\nu}\}=\lambda/(1+\lambda). \tag{18}$$

It then follows immediately, since $(W-M\sqrt{h})/S\sqrt{(1+h)}$ is distributed as t with $\nu+1$ degrees of freedom, that

$$k=\sqrt{\{(1+h)\nu/(\nu+1)\}}\,t(\nu+1;\ \lambda/(1+\lambda)\}. \tag{19}$$

Thus the upper normal linear-utility tolerance limit is $m+ks$, where k is given by (19).

There is here no relation between q and λ (independent of ν) which leads to equality of expressions (12) and (19); the complete equivalence of expected-cover and linear-utility intervals, which was a feature of the Bayesian approach, is thus absent. When, however, the estimate s of σ is based on a large sample, so that ν is appreciable, we see from (12) and (19) that the two intervals are for all practical purposes the same if q and λ are related as in (11). Since the direct interpretation of q is not without difficulty (see Aitchison and Sculthorpe, 1965, p. 477) the new interpretation arising from the correspondence (11) provides a useful alternative view of expected-cover intervals. An interval with expected-cover q is for appreciable ν the same as a linear-utility interval with λ-factor equal to $q/(1-q)$. For example, a statistician who uses a 95 per cent expected-cover interval is behaving in approximately the same way as a statistician who regards the proportional loss caused by outcomes above the limit to be 19 times more serious than that caused by outcomes inside the interval.

3. Extensions and Discussion

While we have developed the theory for intervals of type $(-\infty, r)$ it is clear that a similar development is possible for intervals of type (r, ∞), and that the comments on correspondence between expected-cover and linear-utility intervals remain unchanged. For example, if λ again denotes the ratio of the proportional cost of falling outside to that of falling inside the interval, the frequentist expected-cover and linear-utility intervals are $(m-k_0 s, \infty)$ and $(m-ks, \infty)$, where k_0 and k are given as before by (12) and (19).

It may also be remarked that the Bayesian correspondence and the frequentist approximate correspondence can be extended to the case of finite prediction intervals of the form (r_1, r_2). For example, in the normal case, the symmetric linear y-utility specification

$$V(r_1, r_2, y)=\begin{cases}\lambda(y-r_1) & \text{if } y\leqslant r_1,\\ r_1-y & \text{if } r_1<y\leqslant \tfrac{1}{2}(r_1+r_2),\\ y-r_2 & \text{if } \tfrac{1}{2}(r_1+r_2)<y\leqslant r_2,\\ \lambda(r_2-y) & \text{if } y\geqslant r_2,\end{cases}$$

leads to such a correspondence. It must be admitted, however, that we are not aware of any prediction problem for which such a specification is realistic.

By a development similar to that of Section 2 it can be shown that for the case of gamma distributions, with

$$p_E(x\,|\,\theta)=\theta^k x^{k-1}e^{-\theta x}/\Gamma(k),$$
$$p_F(y\,|\,\theta)=\theta^l y^{l-1}e^{-\theta y}/\Gamma(l),$$

a frequentist linear-utility interval $\{0, \delta(x)\}$ based on (10) is given by

$$\delta(x)=x\frac{1-B\{k+1, l;\ 1/(1+\lambda)\}}{B\{k+1, l;\ 1/(1+\lambda)\}},$$

where $B\{k+1, l; 1/(1+\lambda)\}$ is the $1/(1+\lambda)$ fractile of the beta distribution with density $t^k(1-t)^{l-1}/B(k+1, l)$. Apart from the replacement of k by $k+1$, this interval is the same as the frequentist interval of expected-cover $q = \lambda/(1+\lambda)$; see Aitchison and Sculthorpe (1965, Table 1). When k is large (which will be so, for example, when E is a large number of replicates of F) the difference between the two types of frequentist interval is small, and again the two ways of interpreting such prediction intervals complement each other.

The fact that relation (11) identifies Bayesian, but not frequentist, expected-cover and linear-utility intervals means that any attempt to derive normal frequentist intervals from their Bayesian counterparts with special priors will involve the use of different priors for the two types of interval. It is in fact easily shown from relation (19) and Aitchison and Sculthorpe (1965, Table 1) that the improper prior density $\pi(\mu, \sigma) \propto 1/\sigma$ applied to the Bayesian expected-cover (or linear-utility) interval produces the frequentist expected-cover interval, whereas to produce the frequentist linear-utility interval requires the use of the improper prior $\pi(\mu, \sigma) \propto 1/\sigma^2$. Such a result is of some interest because it warns against the too ready acceptance of $\pi(\mu, \sigma) \propto 1/\sigma$ as the "ignorance" prior which will reproduce frequentist results. A similar feature arises also in the gamma case. It may be that this "discrepancy" between priors (which in the normal case is of the order of a degree of freedom) can be resolved by allowing sequential experimentation in the manner of the recent interesting investigation by Bartholomew (1965), but this is certainly beyond the scope of the present paper.

ACKNOWLEDGEMENT

The author is grateful to the referee for many helpful comments on an earlier draft of this paper..

REFERENCES

AITCHISON, J. (1964), "Bayesian tolerance regions", *J. R. statist. Soc.* B, 26, 161–175.

AITCHISON, J. and SCULTHORPE, DIANE (1965), "Some problems of statistical prediction", *Biometrika*, 52, 469–483.

BARTHOLOMEW, D. J. (1965), "A comparison of some Bayesian and frequentist inferences", *Biometrika*, 52, 19–35.

BOWKER, A. H. and LIEBERMANN, G. J. (1959), *Engineering Statistics*. Englewood Cliffs: Prentice-Hall.

FRASER, D. A. S. and GUTTMAN, I. (1956), "Tolerance regions", *Ann. math. Statist.*, 27, 162–179.

OWEN, D. B. (1963), *Factors for One-sided Tolerance Limits and for Variables Sampling Plans*. Sandia Corporation Monograph SCR-607.

PROSCHAN, E. (1953), "Confidence and tolerance intervals for the normal distribution", *J. Amer. statist. Ass.*, 48, 550–564.

RAIFFA, H. and SCHLAIFER, R. (1961), *Applied Statistical Decision Theory*. Boston: Graduate School of Business Administration, Harvard University.

THATCHER, A. R. (1964), "Relationship between Bayesian and confidence limits for predictions", *J. R. statist. Soc.* B, 26, 176–192.

WALLIS, W. A. (1951), "Tolerance limits for linear regression", *Proc. 2nd Berkeley Symp. Math. Statist. and Prob.*, 43–52.

8 SOME PROBLEMS OF DECISION MAKING

If problems of prediction can be formulated in decision-making terms then the confrontation of a specified value function $V(R,y)$ with an experimentally based $p(y|x)$ involves straightforward maximisation, leading at best to an explicit solution and at worst to computation towards a numerical result. In engineering situations there is usually some hope of specifying $V(R,y)$ in a realistic way, and even when such a specification leaves some element of vagueness it may be possible to show the range of solutions produced by this vagueness, and so present a basis for evaluation of designs. For an example of this sensitivity type of analysis see Aitchison and Sculthorpe (9:1965).

There are, however, areas of possible application where the specification of utility structures is much more difficult, for example, in clinical medicine. In clinical practice decisions are made in the face of uncertainty, often great uncertainty, and so, however much clinicians may back away from the idea, there must be implicit utility judgments being made about the quality, or even the value, of individual lives. Moreover, if decision making is coherent and rational (we do not need to spell out the precise meaning of these terms in this commentary) then actions are chosen *as if* there is a well-defined utility structure and a probability distribution with actions chosen by maximising expected utility. If clinicians are unable or unwilling to specify their utility structures an interesting question that then emerges is the extent to which it is possible to reconstruct or estimate the implicit utility function from a series of recorded actions.

A suitable area for such an investigation appeared to be in the area of treatment allocation and a feasibility study is described by Aitchison (11:1970). On the basis of information z on a patient (the initial state) the clinician assigns treatment t (one chosen from several available) and the patient's 'final state' is in terms of features y. The predictive aspect of the problem is seen as determining a reasonable assessment of the prognostic distribution $p(y|t,z)$, depicting the probability that a patient in initial state z and given treatment t will display final state y. The feasibility analysis undertaken assumes that the set of prognostic distributions is known or has been constructed. The decision problem is then seen as the specification of a utility structure $V(z,t,y)$ with, for a given patient with initial state z, a choice of t which maximises

$$U(z,t) = \int_Y V(z,t,y)p(y|t,z)dy.$$

The models for inconsistent decision making are then studied reflecting the quality of data which the clinician may provide when faced with deciding on treatments for a sequence of n patients with initial states $z_1,\ldots,z_n$. Even with simple one-dimensional z and y, with $U(z,t)$ taken to be linear in z for each t and with the prognostic distribution $p(y|t,z)$ of normal linear form in z for each t, there are many awkward problems concerning identifiability of parameters. For the poorest quality of data, namely where the clinician can provide no information on any aspect of utilities but only on the chosen treatments $t_1,\ldots,t_n$ for the n patients, the unidentifiability is so extreme that only the breakpoints on the z scale where treatment changes can be estimated.

It would be too much of a digression from our main theme to

detail all the various aspects of this investigation. One interesting technical aspect in relation to many of the estimation problems here is that they fall within the area of probit analysis and its generalisations. One such generalisation is reported in Aitchison and Bennett (12:1970).

AITCHISON, J. (1970)

Statistical problems of treatment allocation

Reprinted from *J. R. Statist. Soc.* A133, 206-28

Statistical Problems of Treatment Allocation

By J. AITCHISON

University of Glasgow

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, January 14th, 1970, the President, Professor SIR ROY ALLEN, in the Chair]

SUMMARY

The problem of allocating one of a number of possible treatments to an individual on the basis of information about his initial state must take account of the associated prognosis distributions for his future state and also the utility structure. A decision-theoretic model for such treatment allocation is presented and within its framework are considered questions of optimality, suboptimality of various kinds, and also the feasibility of recovering the implicit utility structure used by a decision-maker from observation of his pattern of treatment allocation.

INTRODUCTION

THE aims of this paper are modest. The first is to bring to the notice of fellow statisticians a source of interesting and important problems which have had less than their fair share of attention. I am encouraged to believe that such a display is necessary by the reaction of the chairman of a recent university seminar bearing tonight's title. He admitted great surprise at the content of the talk, having expected, or perhaps been conditioned to expect, some new investigation of the design problem of how to allocate treatments to *experimental* units to obtain an efficient comparative trial, and not the problem of day-to-day judgments of how to allocate treatments to *non-experimental* units. The second intention is to demonstrate to workers in other disciplines, and in particular medicine, that their difficult problems of decision-making under uncertainty are under active investigation by statisticians. There are indeed encouraging signs among workers in medicine (see, for example, Card, 1967, 1970; Ledley and Lusted, 1962; Lusted, 1968) of a desire to re-appraise their whole approach to decision-making in more numerate and scientific terms.

By attempting this dual purpose I am well aware that I have probably fallen between two stools. My one fear is that, as a consequence, discussants may spend all their time trying to place me safely back on one, while my one hope is that I may be told how to sit comfortably on both, in short that tonight may provide a forum for the discussion of what may usefully be attempted in this field. May I also apologize in advance for the absence of a detailed application to a real problem. The explanation is simple; I have as yet no suitable real data. Indeed one of the reasons for undertaking this work is the belief that any request for funds to sponsor a real investigation should be preceded by the exploration of the possibilities. My choice therefore seemed to lie between holding back the ideas of the paper for n ($\geqslant 3$) years until such time as data are collected and analysed, or presenting this kind of feasibility study. The hope that others may recognize their problems in tonight's formulation and already have suitable data has made me bold enough to make the first choice.

1. FORMULATION

There are many practical problems which have the following general form. An individual unit is initially in a state x, where x (possibly a vector) belongs to a set X of possible initial states. A set T of alternative treatments is available for application to the individual unit. The application of a chosen treatment brings the individual unit to a final state y, a member of a set Y of final states, but not all units in initial state x and given treatment t will reach the same final state y. For given x and t the variability in y is described by some *prognosis density function* $p(y|x,t)$ on Y. The net benefit of applying treatment t to a unit in initial state x depends on the superiority, in some sense, of y over x, and on the cost, in some sense, of applying the treatment to x; in other words, some *utility function* $U(x,t,y)$ is explicitly, or more commonly implicitly, at the basis of any rational procedure for allocating treatments to units. A treatment allocator, or more briefly an *allocator*, is then simply a function τ on X to T; for each $x \in X$ an allocator τ provides a treatment or treatment *allocation* $\tau(x) \in T$.

A few practical situations will show the general nature of the problem and highlight differences in the availability of information on $p(y|x,t)$ and in the difficulty of specifying $U(x,t,y)$.

Example 1. *Improving process quality.* An attempt is to be made to rationalize the method of allocating treatments (one of which may be "do not treat") to loads of raw material of differing initial quality x to obtain a final quality y. Past experience may be so extensive and the allocation of treatment to initial quality x so haphazard that a full picture of the $p(y|x,t)$ density functions is available from records. Moreover we may know the cost k_t of treatment t for each t, and also the way in which selling price $g(x)$ depends on the quality x at which a load is marketed. Thus the utility function is completely specified as

$$U(x,t,y) = g(y) - g(x) - k_t. \tag{1.1}$$

This is a situation where there is perfect information with completely specified prognosis and utility functions.

Example 2. *Initial state an indicator of final quality.* There is nothing in the formulation which demands that x and y should be of the same nature, that X and Y should be the same set. For example, x may be some indicator (present height, degree of pest infestation, the extent of weeds) of potential yield of a growing crop; the treatments the different combinations of fertilizer, insecticide and weeding practice; and y the eventual yield. Here $U(x,t,y)$ may again be specified, a reasonable form being

$$U(x,t,y) = g(y) - k_t, \tag{1.2}$$

where $g(y)$ is the market price of a crop of yield y and k_t is the cost of treatment t. To obtain information about $p(y|x,t)$ in this case it is probably necessary to conduct a controlled experiment or field trial, in which crops at different indicator levels $x_1, \ldots, x_n$ are assigned treatments $t_1, \ldots, t_n$ in some specified random way and the resulting yields $y_1, \ldots, y_n$ are determined. The informative experiment thus provides data in the form of n triplets $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$.

One problem of which I have direct experience and which lies somewhere between Examples 1 and 2 is that of optimum selection of electricity tariffs for hospitals; see the report by Thomson (1968), who does not go fully into the decision-theoretic approach now briefly described. Electricity Boards offer a choice of at least three tariffs, the bill at the end of a basic period depending on total consumption x, the

"treatment" or tariff t chosen, and peak instantaneous demand y during the period. Careful study of the tariff conditions soon shows that the amount of the bill or "loss" $L(x, t, y)$ can be expressed, for each t, as a function which is piecewise linear within simple subsets of the (x, y) plane. A prospective customer such as a hospital may have a shrewd idea of what range of x's it will operate at, but have little idea of what the more nebulous quantity y will be. Studies by Thomson of data on (x, y) from a number of hospitals extending over a period of years suggest that $p(y|x, t)$ takes simple linear normal regression forms. He in fact makes the added assumption that t does not affect y. While this is probably reasonable for the complex behaviour of a hospital, for other users there may be a "treatment effect", the choice of a t which punishes large y inducing a conscious attempt by the user to keep y low. The confrontation of $L(x, t, y)$ with $p(y|x, t)$ by the method of the next section then produces the optimum tariff appropriate to any particular indicator x.

Example 3. *Allocation of a medical treatment.* A patient consults a physician because there are some unpleasant features in his present state or symptom vector x. Here x will consist of information such as high temperature, backache, dizziness, and may be added to by the physician when he records, for example, pulse rate, blood pressure, result of a blood test. The physician may well reach his decision about which treatment to assign to the patient in two stages. First as an aid to his thinking he *diagnoses* the ailment. He says that the cause of the abnormalities in the symptom vector is the presence of some disease, or at least one of a restricted class of diseases. (The statistical problem here is that of *classification*, and will not be our immediate concern although it will arise in Section 6.7 in an unusual setting. Classification has of course been a recent subject of discussion at our Society; see, for example, Hills (1966), Marshall and Olkin (1968).) Having completed the diagnosis stage and so delimited his problem, he then begins to think of what treatment—medication by one of a number of available drugs, or surgery—he should assign to the patient in order to bring the present symptom vector x to some final symptom vector y which he hopes will be more pleasant than x for the patient. In his assignment of treatment he must take account of the variability of prognosis even with a given treatment, the desirability of moving from unpleasant x to more pleasant y, and the differing "costs" of treatments, not only in money terms such as occupation of hospital bed, commitment of nursing staff, drug bill, but as importantly in terms of discomfort of treatment to the patient.

The source of the information on which the physician bases his prognosis distributions $p(y|x, t)$ is a combination of his training, his reading and his experience. It is a remarkable phenomenon that this distribution is usually stored in the physician's head and is seldom set down in any concrete form. Again the utility structure is seldom, if ever, thought of in any explicit way. And yet treatments are allocated, decisions are made. Can any constructive analysis be undertaken in such situations to relieve the physician of some of the burden of storing information and to direct his attention more directly to the precise nature of the value judgments he is implicitly making?

The only references I have been able to find to a decision-theoretic approach to treatment allocation in medicine are Raiffa (1968), who in his introductory "thumbnail sketches" presents and discusses in general terms on pp. 250–255 the decision problem of treatment of a sore throat caused by streptococci or a virus; and Lusted (1968, pp. 150–159), whose main analysis is in programming terms; see also Aitchison (1970b). Most writers seem preoccupied with diagnosis.

In discussions with medical and dental colleagues the following areas have been suggested as possibly suitable for investigation.

(i) *The operation of a coronary care unit.* Here the symptom vector, in addition to "medical" data, would contain information on age and family responsibility; treatments may be simplified to confinement to the intensive care unit or to the general ward; the final state may be some measure of mobility at three or six months' time.

(ii) *Treatment of thyrotoxicosis.* Statistical studies of diagnosis in thyroid clinics are among the most successful and detailed yet undertaken; see, for example, Boyle *et al.* (1966) and Taylor (1970). Experience, goodwill, and continuing studies by Dr Taylor indicate that the considerable controversy over the relative merits of treatments for thyrotoxicosis—surgery, orally administered radioactive iodine, and drugs—would be an interesting choice for study.

(iii) *Root treatment in dentistry.* In endodontics the "hollow tube" controversy (Kennedy and Simpson, 1969) is concerned with the best choice of treatment for root infection of a tooth—extraction, the conventional "endodontic triad" (criticized by Seltzer and Bender, 1965) of thorough débridement, sterilization and complete obturation of the root canal, a variety of other treatments which delay permanent filling until there is radiographical evidence of healing, and surgery. An attempt to investigate the factors in x with somewhat limited data is to be found in Storms (1969), and discussion of the elements of y in Bender *et al.* (1966) and Seltzer *et al.* (1967). In such studies the "cost" of treatments certainly has to take into account the staying power of the patient over what may turn out to be a necessarily long period of treatment.

Example 4. *Sampling inspection.* Some familiar problems of sampling inspection may be readily expressed within the framework of treatment allocation. In our example here the initial state x is unknown and an informative experiment provides data on which to base some plausibility distribution $p(x|z)$ for x.

A batch of N items from a batch process contains unknown numbers x of defective and $N-x$ of effective items. From past experience we may know the plausibilities $p(x)$ attaching to the various x. The different treatments may consist of inspecting different numbers of items and rectifying any defectives in the inspected set before placing the batch on the market. One method of proceeding is to carry out a preliminary inspection of a random sample of size n from the batch. Suppose that z are found to be defective and $n-z$ effective. Since

$$p(z|x) = \binom{x}{z}\binom{N-x}{n-z}\Big/\binom{N}{n} \tag{1.3}$$

is known, we can convert the prior information $p(x)$ into posterior information $p(x|z)$ by a simple application of Bayes's theorem. The information z is now to be used to decide how many additional items t should be inspected before placing the batch on the market. The cost of inspecting each item is k; all detected defective items are rectified at cost q per item; the replacement loss of a defective item is r; the selling profit of an effective item is s.

Here y is the eventual number of defectives in the batch and it is clearly necessary to write the prognosis distribution as $p(y|x,z,t)$ to display its dependence on z as well as on x and t. In this case we can completely specify

$$p(y|x,z,t) = \binom{x-z}{x-z-y}\binom{N-x-n+z}{t-x+z+y}\Big/\binom{N-n}{t}. \tag{1.4}$$

Also a suitable utility function can be constructed:

$$U(x, z, t, y) = Ns - ry - q(x-y) - c(z+t), \tag{1.5}$$

since we sell N items, have to replace y, have rectified $x-y$ and have inspected altogether $z+t$.

To resolve the allocation problem we have to confront the utility $U(x, z, t, y)$ with the combined plausibility and prognosis assessment $p(x|z)p(y|x, z, t)$.

Example 5. *An economic situation.* An extremely difficult form of the problem is that facing the Chancellor of the Exchequer on budget day. Here x is the present economic state of the country, the treatments are the possible budgets he may present, and y is the economic state of the country at some specified future date or dates. The problem is difficult for two main reasons, an inherent unwillingness even among political economists of the same affiliation to express with any clarity their $U(x, t, y)$ and the almost total absence of experience on which to base a sound prognosis distribution $p(y|x, t)$.

From these five examples we see that the problem presents itself in varying degrees of completeness, from the case of a known utility structure and prognosis distribution in Example 1; through the case of a known utility structure with a prognosis distribution to be inferred from the results of a controlled experiment in Example 2; the lack of formal expression of a utility function and the replacement of a specific prognosis analysis by the use of informal experience in Example 3; the lack of precise knowledge of the present state in Example 4; to the precarious problem of Example 5 where vague aims and vague prognostications abound in a situation where it is impossible to experiment.

It is the purpose of this paper to investigate and illustrate what can be effectively achieved in situations of different degrees of clarity about aims and of information about the effects of treatments. In Section 2 the optimum theory under perfect information on $p(y|x, t)$ and $U(x, t, y)$ is first set out as a yardstick, and in Section 3 a normal linear utility model, used to illustrate concepts and results, is briefly described. Measures of suboptimality associated with a number of suboptimal allocators of interest are obtained in Section 4. The use of an informative experiment to provide information on $p(y|x, t)$ is described in Section 5, only briefly since our main interest in this paper is in situations where the main cause of difficulty is the reluctance to specify the utility function. In many of these vaguely specified practical situations treatments are actually allocated by some decision-maker.

A question of considerable interest will be that of attempting to describe such actual treatment allocation in terms of statistical decision theory and the particular problem of discovering or recovering the utility function the decision-maker is subconsciously using. For this it is necessary to introduce a model of an inconsistent decision-maker. The extent to which his utility function is recoverable will depend on the ease with which we can make some suitable assumption about its parametric structure, and the amount of information about his "optimum" decisions that the decision-maker can give us. Different situations are dealt with in Section 6, and recovery estimation methods investigated and illustrated. One of the wilder hopes of such an investigation is that the confrontation of the decision-maker with his utility structure may eventually bring to him an awareness of the possibility of formalizing his decision processes and perhaps relieve him of the wearing necessity of much of the more routine decision-making he has repeatedly to face.

On the whole the discussion is confined to the case of two treatments, although many of the results extend to the case of more than two treatments. The extension, however, is not a routine matter and involves in particular some interesting problems of identifiability and estimability of parameters. In the fear that these may obscure the structure of the problem I have not considered them in this paper.

2. The Case of Perfect Information

The case where the initial state x, the prognosis density functions $p(y|x,t)$ and the utility function $U(x,t,y)$ are all known to the decision-maker is straightforward. Our interest in this case of perfect information is not only for its own sake, but also as a basis for subsequent considerations. A standard principle on which to obtain an *optimum allocator* τ^* is then to ensure that it allocates treatments so as to maximize the expected utility

$$U(x,t) = \int_Y U(x,t,y)p(y|x,t)\,dy, \tag{2.1}$$

where the expectation operation is with respect to the prognosis distribution for given (x,t). (Throughout theoretical considerations, we shall assume that y is a continuous random variable since our illustrative example takes this form; the case of discrete random variables, or random vectors, possibly of mixed type, requires only the replacement of the simple integration operation by the appropriate accumulation operation.) An optimum allocator can thus be constructed by choosing, for each $x \in X$, a treatment $\tau^*(x)$ satisfying

$$U\{x,\tau^*(x)\} = \max\{U(x,t): t \in T\}. \tag{2.2}$$

It may happen that $\tau^*(x)$ is constant, say t^*, for all $x \in X$, that is, the same treatment t^* is optimum for every initial state. Such a treatment t^* is termed a *uniformly optimum allocation.*

Note that in the case of perfect information no informative experiment on which to base prognosis assessments is necessary. Another point worth noticing at this stage is the special feature that the action or decision t influences the unknown "state of nature" y. It is in fact a problem in predictive decision-making in the sense of Aitchison and Sculthorpe (1964).

3. Normal Linear Utility Model

Since it is likely to play an important role in subsequent theoretical development and first attempts at practical application, and also for its usefulness in illustrating the theory and concepts, we specify, and establish the notation of, a normal linear utility model. For this model in its n-dimensional multivariate form we have

$$p(y|x,t): \quad \text{a } N(\alpha_t + B_t x, \Sigma) \text{ density function,} \tag{3.1}$$

$$U(x,t,y) = \xi' x + \eta' y - \kappa_t, \tag{3.2}$$

where α_t, ξ, η are vectors and B_t, Σ matrices of appropriate dimensions. Then

$$U(x,t) = (\xi' + \eta' B_t)x + \eta'\alpha_t - \kappa_t. \tag{3.3}$$

In its simple univariate form we have

$$p(y|x,t): \quad \text{a } N(\alpha_t+\beta_t x, \sigma^2) \text{ density function,} \tag{3.4}$$

$$U(x,t,y) = \xi x+\eta y-\kappa_t, \tag{3.5}$$

where all the symbols are scalars.

Two useful integrals $I(a,b,c,d)$ and $J(a,b,c,d)$ associated with the normal model are set out in the Appendix.

4. The Measurement of Suboptimality

4.1. *Definition*

If, from application to application of an allocator τ, the initial state x is variable, with density function $p(x)$ on X, then we obtain an overall measure of the effectiveness of τ in the expected utility per allocation by τ, namely

$$U(\tau) = \int_X U\{x,\tau(x)\}p(x)\,dx. \tag{4.1}$$

The extent of suboptimality of a non-optimum procedure τ can then be readily measured as the loss of expected utility per allocation when compared with an optimum allocator τ^*:

$$S(\tau,\tau^*) = U(\tau^*)-U(\tau). \tag{4.2}$$

The importance of suboptimality in this paper arises mainly in circumstances where we envisage the decision-maker unable or unwilling to enunciate his problem explicitly and as a result indulging in inconsistencies of some kind. Suboptimality provides a measure of the consequences of these inconsistencies.

4.2. *Suboptimality from Use of Wrong Utility Function*

For the case of two treatments with $T=\{1,2\}$ we have the following optimum allocator:

$$\tau^*(x) = \begin{cases} 1 & \text{if} \quad U(x,1)-U(x,2)\geqslant 0, \\ 2 & \text{if} \quad U(x,1)-U(x,2)<0. \end{cases} \tag{4.3}$$

Then

$$\begin{aligned} U(\tau^*) &= \int_{U(x,1)-U(x,2)\geqslant 0} U(x,1)p(x)\,dx + \int_{U(x,1)-U(x,2)<0} U(x,2)p(x)\,dx \\ &= \int_X U(x,2)p(x)\,dx + \int_{U(x,1)-U(x,2)\geqslant 0} \{U(x,1)-U(x,2)\}p(x)\,dx. \end{aligned} \tag{4.4}$$

Suppose that $U(x,t)$ is linear in x, say

$$U(x,t) = a_t+b_t'x, \tag{4.5}$$

then

$$U(x,1)-U(x,2) = a+b'x,$$

where $a=a_1-a_2$, $b=b_1-b_2$. The use of a suboptimal procedure may arise from a wrong assignment of the parameters of the utility structure and so lead to an allocator:

$$\tau(x) = \begin{cases} 1 & \text{if} \quad c+d'x\geqslant 0, \\ 2 & \text{if} \quad c+d'x<0. \end{cases} \tag{4.6}$$

Then

$$U(\tau) = \int_X U(x,2)\,p(x)\,dx + \int_{c+d'x\geqslant 0}\{U(x,1)-U(x,2)\}\,p(x)\,dx$$

so that

$$U(\tau^*)-U(\tau) = \left(\int_{a+b'x\geqslant 0} - \int_{c+d'x\geqslant 0}\right)(a+b'x)\,p(x)\,dx$$
$$= I(a,b,a,b)-I(a,b,c,d), \tag{4.7}$$

from Integral 1 of the Appendix, if x is $N(\psi,\Omega)$.

4.3. *Suboptimality of a Randomized Allocator*

We shall find in later sections that in order to describe the inconsistent behaviour of actual decision-makers we have to allow randomized allocators. For the case where T consists of two treatments such an allocator can be expressed in the form

$$\tau(x) = \begin{cases} 1 & \text{with probability } \pi(x), \\ 2 & \text{with probability } 1-\pi(x). \end{cases} \tag{4.8}$$

Then $U(\tau)$ is easily evaluated as

$$U(\tau) = \int_X U(x,1)\,\pi(x)\,p(x)\,dx + \int_X U(x,2)\{1-\pi(x)\}\,p(x)\,dx$$
$$= \int_X U(x,1)\,p(x)\,dx + \int_X \{U(x,1)-U(x,2)\}\,\pi(x)\,p(x)\,dx.$$

Hence

$$S(\tau,\tau^*) = \int_{U(x,1)-U(x,2)>0}\{U(x,1)-U(x,2)\}\,p(x)\,dx - \int_X \{U(x,1)-U(x,2)\}\,\pi(x)\,p(x)\,dx. \tag{4.9}$$

For one of the normal linear utility models for inconsistent behaviour studied later we shall find that $\pi(x)$ takes the form $\Phi(c+d'x)$. For this normal linear case then we have immediately from Integrals 1 and 2 of the Appendix that the suboptimality of the randomized allocator is

$$I(a,b,c,d)-J(a,b,c,d). \tag{4.10}$$

4.4. *Suboptimality of a Constant Allocator*

Since a constant allocator, say τ with $\tau(x) = 2$ for every $x \in X$, can be expressed as a randomized allocator with $\pi(x) = 0$ for every $x \in X$, the above result allows us to assess the suboptimality of misassuming that treatment 2 is a uniformly optimum allocation. For such a constant allocator the suboptimality reduces to

$$S(2,\tau^*) = \int_{U(x,1)-U(x,2)>0}\{U(x,1)-U(x,2)\}\,p(x)\,dx. \tag{4.11}$$

Such a measure is relevant, for example, when it is mistakenly assumed in the investigation of a controlled clinical trial with two treatments that the only conclusions possible are that both treatments are equally effective or that one treatment is better than the other *for all initial states.*

4.5. *Suboptimality from Use of Subvector of Initial State Vector*

The motivation for considering this problem is that it arises whenever a decision is required as to whether to allocate a treatment on the basis of a subvector x_1 of x or whether first to gain additional information by observing the remainder x_2 at some cost and then to allocate on the basis of the complete vector (x_1, x_2). Analysis of this situation is a first step towards a full sequential treatment of the problem. In addition to the prognosis distribution $p(y|x_1, x_2, t)$ and the utility specification $U(x_1, x_2, t, y)$ we assume that the conditional density function $p(x_2|x_1)$ is known also. Clearly the variability of x_2 for given x_1 is very relevant to the decision whether or not to allocate on the basis of x_1 alone.

A convenient way of expressing the advantages or disadvantages of taking the additional information is to quote a *breakeven cost*, that is the maximum amount that it would be worth paying for the additional information. This is equivalent to the concept of the expected value of information of Raiffa and Schlaifer (1961, p. 87).

Let us consider first allocators based on x_1 alone, that is functions from X_1 to T. The unknowns in the utility $U(x_1, x_2, t, y)$ are (x_2, y) and so the allocation of treatment at this stage (that is, on information x_1 only) is to be assessed in terms of the expected utility with respect to $p(y|x_1, x_2, t)p(x_2|x_1)$, that is,

$$U(x_1, t) = \int_{X_2 \times Y} U(x_1, x_2, t, y)\, p(y|x_1, x_2, t)\, p(x_2|x_1)\, dy\, dx_2$$

$$= \int_{X_2} U(x_1, x_2, t)\, p(x_2|x_1)\, dx_2.$$

The optimum allocator τ_1^* thus assigns allocation $\tau_1^*(x_1)$ so as to maximize this. The maximum expected utility from an immediate allocation on the data x_1 is therefore

$$\max_T \int_{X_2} U(x_1, x_2, t)\, p(x_2|x_1)\, dx_2.$$

To compare this with what we might gain with the taking of additional information x_2 we will assume that the cost of observing x_2 is zero and that $U(x_1, x_2, t, y)$ assesses the advantages and disadvantages of other factors. If we were to observe x_2 then we must consider allocators: $(x_1, x_2) \to \tau(x_1, x_2)$. The optimum allocator τ^* will then provide expected utility

$$\max_T U(x_1, x_2, t).$$

At the point of assessing the advantages of τ^* over the previous τ_1^* we, of course, know only x_1, and so we must take the expectation with respect to $p(x_2|x_1)$. The maximum expected utility obtainable from making use of this cost-free additional information is

$$\int_{X_2} \max_T U(x_1, x_2, t)\, p(x_2|x_1)\, dx_2.$$

Hence the breakeven cost, which is simply a measure of the suboptimality of using information x_1 alone as compared with using cost-free further information, is

$$S(\tau_1^*, \tau^*) = \int_{X_2} \max_T U(x_1, x_2, t)\, p(x_2|x_1)\, dx_2 - \max_T \int_{X_2} U(x_1, x_2, t)\, p(x_2|x_1)\, dx_2. \quad (4.12)$$

Example. For the two-treatment normal linear utility model we set

$$U(x_1, x_2, t, y) = \xi_1' x_1 + \xi_2' x_2 + \eta' y - \kappa_t,$$

$$p(y|x_1, x_2, t) = \phi(y|\alpha_t + B_t x_1 + \Gamma_t x_2, \Sigma),$$

$$p(x_2|x_1) = \phi(x_2|\mu, \Omega).$$

The parameters μ and Ω will naturally depend on x_1 but we have dropped the notational dependence for the sake of simplicity. Then

$$U(x_1, x_2, t) = \xi_1' x_1 + \xi_2' x_2 + \eta'(\alpha_t + B_t x_1 + \Gamma_t x_2) - \kappa_t.$$

Now

$$\tau^*(x_1, x_2) = \begin{cases} 1 & \text{if} \quad U(x_1, x_2, 1) - U(x_1, x_2, 2) \geqslant 0, \\ 2 & \text{if} \quad U(x_1, x_2, 1) - U(x_1, x_2, 2) < 0, \end{cases}$$

and an application of Integral 1 of Section 3 gives

$$\int_{X_2} \max_T U(x_1, x_2, t)\, p(x_2|x_1)\, dx_2 = U(x_1, \mu, 2) + w[\phi\{v(x_1)\} + v(x_1)\, \phi\{v(x_1)\}], \tag{4.13}$$

where

$$w = \{\eta'(\Gamma_1 - \Gamma_2)\, \Omega (\Gamma_1 - \Gamma_2)' \eta\}^{\frac{1}{2}}, \tag{4.15}$$

$$v(x_1) = w^{-1}\{U(x_1, \mu, 1) - U(x_1, \mu, 2)\}. \tag{4.16}$$

Hence the breakeven cost is

$$b(x_1) = \begin{cases} w[\phi\{v(x_1)\} + v(x_1)\, \Phi\{v(x_1)\} - v(x_1)] & \text{if} \quad v(x_1) \geqslant 0, \\ w[\phi\{v(x_1)\} + v(x_1)\, \Phi\{v(x_1)\}] & \text{if} \quad v(x_1) < 0, \end{cases} \tag{4.17}$$

and

$$\max_{X_1} b(x_1) = (2\pi)^{-\frac{1}{2}} w, \tag{4.18}$$

the maximum occurring when $v(x_1) = 0$. Thus if Ψ is defined by the relation (see De Groot, 1968):

$$\Psi(v) = \phi(v) - v\{1 - \Phi(v)\}, \tag{4.19}$$

we reach the following conclusion.

It is worth observing x_2 if the cost involved is less than

$$w\Psi\{v(x_1)\} \quad \text{when} \quad v(x_1) > 0,$$

$$w[\Psi\{v(x_1)\} + v(x_1)] \quad \text{when} \quad v(x_1) < 0. \tag{4.20}$$

If the cost involved exceeds $(2\pi)^{-\frac{1}{2}} w$ then it is never worth observing x_2.

Note that the quantity w is a natural one to enter these cost considerations since we would clearly be prepared to pay more for the additional information the "bigger" η is, the greater the "difference" between Γ_1 and Γ_2, and the more "variability" (Ω) there is in x_2 for given x_1.

5. Estimation of the Prognosis Distribution

For the case where x is known and $U(x, t, y)$ well formulated, but $p(y|x, t)$ not fully known we suppose that the class of prognosis models is indexed by a possibly vector parameter of an index set Θ, $p_\theta(y|x, t)$ being the prognosis density function corresponding to the index $\theta \in \Theta$. For instance in the normal linear case, if there is no great past experience with the treatments, the α_t, B_t and Σ act as indices. A satisfactory informative experiment by which to gain knowledge of θ will then be a controlled trial where individual units with initial states, say $x_1, \ldots, x_n$, providing a good cover of X, are used. The assignment of treatments $t_1, \ldots, t_n$, which may include the treatment "do not treat", to the units in this trial should cover T and be made in a known random way according to the well-established principles of experimental design, the use of blocking or other devices depending on the special circumstances of the situation. The final states $y_1, \ldots, y_n$ are recorded, so that the data then take the form $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$. From a Bayesian viewpoint these data, denoted briefly by z, will convert a prior plausibility function $p(\theta)$ into a posterior plausibility function $p(\theta|z)$, which will then provide a prognosis density

$$p(y|x, t|z) = \int_\Theta p_\theta(y|x, t)\, p(\theta|z)\, d\theta. \tag{5.1}$$

An alternative approach from a frequentist point of view is to obtain an estimate, $\hat{\theta}(z)$ say, of θ by some standard estimation principle, such as maximum likelihood, and set

$$p(y|x, t|z) = p_{\hat{\theta}(z)}(y|x, t). \tag{5.2}$$

It is well known that if the informative experiment is sufficiently large or a suitable indifference form is used for $p(\theta)$ these two prognosis assessments (5.1) and (5.2) will hardly differ. It is certainly not our purpose here to dwell on the relative merits of the two approaches.

Once the prognosis distribution (5.1) or (5.2) has been obtained the treatment allocation problem is again straightforward, the definitions and analysis taking exactly the same form as for the case of perfect information (Sections 2 and 4) with $p(y|x, t|z)$ replacing $p(y|x, t)$.

Example. For the univariate form of the normal linear model with two treatments we suppose

$$\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \sigma)$$

to be unknown; data of the form

$$(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$$

then correspond to two separate regression scatter diagrams with common error variance. Let $\hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2$ be the usual least-squares estimates of α_1, β_1, α_2, β_2, and s^2 the usual "pooled" estimate of σ^2. Then a standard Bayesian inference based on a vague prior $p(\alpha_1, \beta_1, \alpha_2, \beta_2, \sigma) \propto 1/\sigma$ leads to a normal-gamma type of posterior distribution for θ, and it can then be shown that $p(y|x, t|z)$ is of generalized Student form with mean vector $(\hat{\alpha}_1 + \hat{\beta}_1 x, \hat{\alpha}_2 + \hat{\beta}_2 x)$. Subsequent analysis of the treatment allocation problem then follows from the previous considerations of the normal linear model with $\hat{\alpha}_t, \hat{\beta}_t$ replacing α_t, β_t. The frequentist form for $p(y|x, t|z)$, given by (5.2), is $N(\hat{\alpha}_t + \hat{\beta}_t x, s^2)$ and use of this in place of the previous $p(y|x, t)$ again leads to the same subsequent analysis as the Bayesian approach.

This brief and somewhat superficial account of the construction of the prognosis distribution is not an attempt to write the problem off as either trivial or unimportant. Indeed the contrary is the case. The non-triviality of many of the statistical problems is soon realized in some applications, where the final states are (possibly multi-dimensional) categories, for example, death, survival with recurring headache and dizziness, survival with recurring headache but no dizziness, and so on. In such circumstances some interesting multidimensional generalizations of quantal analysis can arise; see Aitchison and Bennett (1970). There is also a growing awareness of the importance of obtaining reliable prognosis distributions in a number of fields, particularly in medicine; see, for example, Peel *et al.* (1962), Hughes *et al.* (1963), Ginsberg and Offensend (1968), Norris *et al.* (1969), and the report of a proposed prognosis study of the treatment of breast cancer in *The Scotsman* of November 21st, 1969. Our reason for not treating this aspect in greater depth in this paper is that we want to concentrate attention on certain problems in the specification of the utility structure. In what follows therefore we shall suppose that the prognosis distribution $p(y|x,t)$, whether constructed along the lines of this section or not, is available.

6. Models for Inconsistent Treatment Allocation Utility Estimation

6.1. *The Problem*

So far we have been concerned with situations where the utility function $U(x,t,y)$ is completely specified, and where the problem is to determine the optimum allocator. Now we turn to the converse problem, where allocations, presumably optimum in the view of the decision-maker, are being made and yet the decision-maker is unwilling or unable to state explicitly his utility function, though he may be prepared to make some utility assessments about the *particular* situations he meets. Can we then infer or estimate the utility function that the decision-maker is implicitly using? The motivation here is to show the decision-maker that it may be possible to formalize his allocation procedure. There are two possible effects of this. The exposure of his utility function may invite him to ask the question: Is this the kind of utility function I want to use, should I not perhaps modify my utility function? Alternatively he may be encouraged to use the estimated utility function to automate the decision-making process and so free himself from much of the arduous routine allocation that comes his way.

The method of recovery of the utility function will depend on what information is supplied by the decision-maker. We shall consider a number of different qualities of information in decreasing order of quality and the different models which generate these data. We would expect a decision-maker who works with a vague utility function not to be consistent in all his allocations, and our models will have to entertain this possibility. We shall thus be constructing models for inconsistent decision-making. One of the topics of interest will then be to measure the extent of this inconsistency, which again we shall interpret in terms of suboptimality. The general type of model considered here we term an *expectation* model. Another type, a *prediction* model, is briefly discussed in Section 6.8.

In order to recover the utility function we shall have to impose some structure on the situation. For the reasons indicated in Section 5 we assume that we know the prognosis density function $p(y|x,t)$. We shall also assume that $U(x,t,y)$ is of a certain parametric form, say $U(x,t,y,\zeta)$, where ζ is the indexing parameter. For example, with the univariate normal linear utility model discussed in Section 3 we may regard

$\zeta = (\xi, \eta, \kappa_1, \kappa_2)$ as the indexing parameter, or if the costs of treatments are known, $\zeta = (\xi, \eta)$. We write

$$U(x, t, \zeta) = \int_{\gamma} U(x, t, y, \zeta) p(y|x, t)\, dy \tag{6.1}$$

for the expected utility. Then for the two-treatment case we have

$$U(x, t, \zeta) = \begin{cases} \eta\alpha_1 - \kappa_1 + (\xi + \eta\beta_1)x & \text{for } t = 1, \\ \eta\alpha_2 - \kappa_2 + (\xi + \eta\beta_2)x & \text{for } t = 2, \end{cases} \tag{6.2}$$

$$= \begin{cases} \lambda_1 + \mu_1 x & \text{for } t = 1, \\ \lambda_2 + \mu_2 x & \text{for } t = 2, \end{cases} \tag{6.3}$$

where the two alternative sets $(\xi, \eta, \kappa_1, \kappa_2)$ and $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ of parameters are in one-to-one correspondence by

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 & \alpha_1 & -1 & 0 \\ 0 & \alpha_2 & 0 & -1 \\ 1 & \beta_1 & 0 & 0 \\ 1 & \beta_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \\ \kappa_1 \\ \kappa_2 \end{bmatrix}, \tag{6.4}$$

$$\begin{bmatrix} \xi \\ \eta \\ \kappa_1 \\ \kappa_2 \end{bmatrix} = \frac{1}{\beta_1 - \beta_2} \begin{bmatrix} 0 & 0 & -\beta_2 & \beta_1 \\ 0 & 0 & 1 & -1 \\ \beta_2 - \beta_1 & 0 & \alpha_1 & -\alpha_1 \\ 0 & \beta_1 - \beta_2 & \alpha_2 & -\alpha_2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{bmatrix}. \tag{6.5}$$

The problem is then whether it is possible to estimate $(\lambda_1, \lambda_2, \kappa_1, \kappa_2)$ or equivalently $(\xi, \eta, \kappa_1, \kappa_2)$.

I can give no general guidance as to choice of parametric form. As in all applied mathematics, simplicity and tractability will largely outweigh other considerations in the early stages of development. For the linear form, here chosen for simplicity, with ξ and η of opposite signs, $U(x, t, y)$ gives some measure of the "difference" between x and y, discounted by cost.

6.2. *Classification of Expectation Models*

When faced with an individual in initial state x the decision-maker may visualize the consequences of applying treatment t through his knowledge (whether formal or informal) of the prognosis distribution and from this assess, possibly subconsciously, the expected utility $U(x, t, \zeta)$. We can then picture him as running mentally through all the possible treatments and obtaining estimates of all the $U(x, t, \zeta)$, estimates subject to errors e_t because of the inherent informality of the process. Thus he arrives at a set of estimated utilities

$$u_t = U(x, t, \zeta) + e_t \quad (t \in T). \tag{6.6}$$

Since we have subsumed in our formulation of the treatment allocation problem that treatments are defined in such a way that two treatments cannot be applied to the same individual we may reasonably assume that our decision-maker produces errors e_t that are independent. We assume moreover that each e_t is $N(0, \delta^2)$. The assumption of homoscedastic variance here would be reasonable, for example, if the decision-maker has much the same experience with all the treatments. We have made the simplest possible assumptions; these are not in any way essential to the arguments that follow though naturally they simplify the mathematical development.

We can then visualize three main possibilities about the kind of information suppliable by the decision-maker.

(1) The decision-maker may be able to provide for each of the presented initial states $x_1, \ldots, x_n$ his expected utility estimates $u_{11}, \ldots, u_{1k}; u_{21}, \ldots, u_{2k}; \ldots; u_{n1}, \ldots, u_{nk}$, for each of the k treatments. We shall term the model which generates data of this form the *full utility expectation model.*

(2) In many situations it will be asking too much of the decision-maker to provide estimated utilities for all the treatments for each of the initial states presented to him. He may be more willing to provide such utilities only for the treatments that he would actually apply for the given initial state, that is, for the treatment that he regards as the appropriate optimum. Thus, if he is presented with a set $x_1, \ldots, x_n$ of initial states which provide a good cover of X the data will take the form

$$(x_1, t_1^*, u_1^*), \ldots, (x_n, t_n^*, u_n^*),$$

where t_i^* denotes the treatment he regards as optimum for the initial state x_i and u_i^* is the utility he associates with this allocation. The model here is termed the *optimum utility expectation model.*

(3) He may feel completely unable to provide any information on the expected utilities but only name the optimum treatments $t_1^*, \ldots, t_n^*$ that he associates with the initial states $x_1, \ldots, x_n$. The corresponding model is the *optimum treatment expectation model.*

6.3. *Data for Illustrative Example*

To illustrate the degree of success of the recovery operation by estimation procedures associated with the three expectation models we have simulated a decision-maker who allocates treatments from $T = \{1, 2\}$ with a univariate normal linear utility form. Indexing in terms of $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ gives the simpler estimation procedures. The following are the values chosen for simulation purposes:

$$\lambda_1 = 7, \quad \lambda_2 = 3, \quad \mu_1 = 0{\cdot}4, \quad \mu_2 = 0{\cdot}8, \quad \delta = 0{\cdot}4.$$

The errors were easily generated by taking $N(0, 1)$ random deviates and multiplying each by 0·4. The utility and treatment data associated with 25 initial states are presented in Table 1.

In this simple example the decision-maker would of course have to be very naive not to recognize his inconsistency in the overlap of 1's and 2's in the column of t^*. I hope the reader will admit this naivety in such an illustrative example. In less orderly sets X in higher dimensions recognition of inconsistencies requires great sophistication. Inconsistencies occur in diagnosis and no doubt also in treatment allocation.

TABLE 1

Simulated data from the three expectation models

Initial state x_1	*Full utility model*		*Optimum utility model*		*Optimum treatment model*
	u_{1i}	u_{2i}	t_i^*	u_i^*	t_i^*
3·1	8·17	5·32	1	8·17	1
4·6	9·01	7·26	1	9·01	1
5·2	9·16	7·38	1	9·16	1
6·4	8·92	8·49	1	8·92	1
6·7	10·37	8·97	1	10·37	1
7·7	9·96	9·51	1	9·96	1
8·7	10·31	10·17	1	10·31	1
9·0	10·29	10·55	2	10·55	2
9·1	10·90	10·07	1	10·90	1
9·5	10·67	10·96	2	10·96	2
9·5	11·04	10·45	1	11·04	1
10·6	11·17	11·42	2	11·42	2
10·9	12·39	11·45	1	12·39	1
11·3	11·08	12·60	2	12·60	2
11·9	12·41	12·70	2	12·70	2
11·9	11·73	12·18	2	12·18	2
13·9	11·78	14·55	2	14·55	2
15·2	13·26	14·80	2	14·80	2
17·7	14·62	17·37	2	17·37	2
19·2	15·00	18·47	2	18·47	2
19·6	15·09	19·22	2	19·22	2
20·9	15·12	18·77	2	18·77	2
22·3	16·09	21·47	2	21·47	2
23·4	16·69	21·51	2	21·51	2
25·0	16·74	22·95	2	22·95	2

6.4. *Full Utility Expectation Model*

With such full data the estimation problem is one of standard regression analysis, the two sets

$$(x_i, u_{1i}) \quad (i = 1, \ldots, n),$$

$$(x_i, u_{2i}) \quad (i = 1, \ldots, n),$$

being the scatter data for two regression lines $u = \lambda_1 + \mu_1 x$, $u = \lambda_2 + \mu_2 x$. Standard least-squares estimation then gives as estimates of $\lambda_1, \lambda_2, \mu_1, \mu_2$,

$$l_1 = 7{\cdot}01, \quad l_2 = 3{\cdot}30, \quad m_1 = 0{\cdot}405, \quad m_2 = 0{\cdot}785;$$

and for the error standard deviation δ the usual root mean-square error (pooling the residual sums of squares of the two regression analyses)

$$d = 0{\cdot}404.$$

6.5. *Optimum Utility Expectation Model*

The data are here displayed in Fig. 1, and, in terms of regression analysis, we can imagine their generation as follows. At each of a number of values of the explanatory variable x, two regression experiments 1 and 2, corresponding to the two treatments and with true regression functions $\lambda_1+\mu_1 x$ and $\lambda_2+\mu_2 x$, are performed. What is recorded, however, is not both responses, but only the greater, together with the number of the experiment, 1 or 2, from which it came. Although Fig. 1 has the

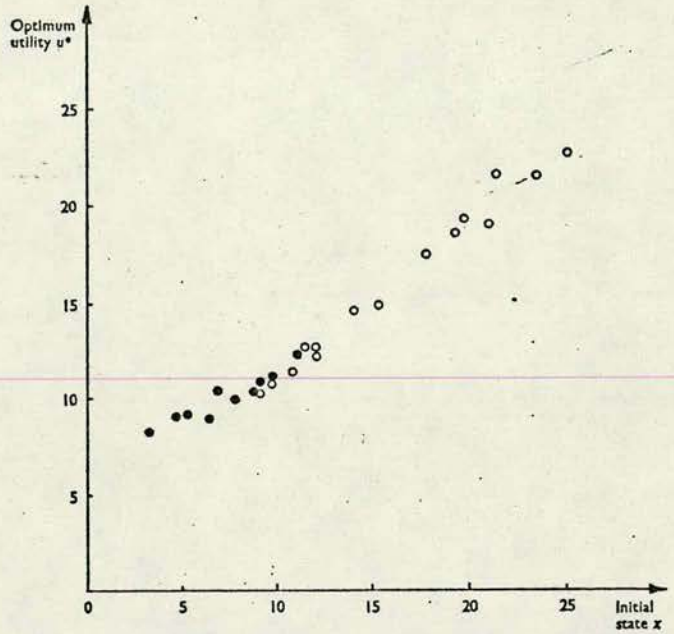

FIG. 1. The data of the optimum utility expectation model of Table 1. Treatment 1, ●; treatment 2, ○.

appearance of a straightforward two regression-line scatter diagram, we are therefore aware that corresponding to each ● there is a ghost ○ below and to each ○ there is a ghost ● below. While it is tempting to fit least-squares regression lines to the two sets of data this is an entirely incorrect estimation procedure. The nature of the generation of the data—a form of random censoring in that the smaller of the two response outcomes is suppressed—must be taken into account, and will often have the effect of altering appreciably the estimated regression functions.

The estimation can be easily carried out by an iterative maximum-likelihood procedure expressible in terms of repeated adjustments to least-squares estimates. The probability that experiment t^* gives an outcome u^* and that this is greater than the outcome of the independent experiment t is

$$\frac{1}{\delta}\phi\left(\frac{u^*-\lambda_{t^*}-\mu_{t^*}x}{\delta}\right)\Phi\left(\frac{u^*-\lambda_t-\mu_t x}{\delta}\right), \tag{6.7}$$

where we use unstarred t to denote the non-optimum experiment or treatment. If we denote by $\sum_i$ summation over the set of n_i observations for which treatment i gives an optimum then the log-likelihood may be expressed as

$$-n\log\delta-\frac{1}{2\delta^2}\sum_1(u^*-\lambda_1-\mu_1 x)^2+\sum_1\log\Phi\left(\frac{u^*-\lambda_2-\mu_2 x}{\delta}\right)$$

$$-\frac{1}{2\delta^2}\sum_2(u^*-\lambda_2-\mu_2 x)^2+\sum_2\log\Phi\left(\frac{u^*-\lambda_1-\mu_1 x}{\delta}\right). \tag{6.8}$$

The likelihood derivative equations can, after some tedious algebra, be converted into a simple iterative scheme for the computation of the maximum likelihood estimates l_1, m_1, l_2, m_2, d of $\lambda_1, \mu_1, \lambda_2, \mu_2, \delta$. This iterative procedure seems simpler than the standard one based on the Newton method and the information matrix. Let the νth iterates be denoted by the $l_1^{(\nu)}, m_1^{(\nu)}, l_2^{(\nu)}, m_2^{(\nu)}, d^{(\nu)}$, and the estimates obtained from the data regarded as simple scatter diagrams for two regression lines by $l_1, \dot{m}_1, l_2, \dot{m}_2, d$. The function w is defined by

$$w(z)=\frac{\phi(z)}{\Phi(z)}. \tag{6.9}$$

The iterative scheme is then:

$$\begin{bmatrix} l_1^{(\nu)} \\ m_1^{(\nu)} \end{bmatrix} = \begin{bmatrix} l_1 \\ \dot{m}_1 \end{bmatrix} - d^{(\nu-1)} \begin{bmatrix} n_1 & \sum_1 x \\ \sum_1 x & \sum_1 x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_2 w\left(\dfrac{u^*-l_1^{(\nu-1)}-m_1^{(\nu-1)}x}{d^{(\nu-1)}}\right) \\ \sum_2 xw\left(\dfrac{u^*-l_1^{(\nu-1)}-m_1^{(\nu-1)}x}{d^{(\nu-1)}}\right) \end{bmatrix}, \tag{6.10}$$

$$\begin{bmatrix} l_2^{(\nu)} \\ m_2^{(\nu)} \end{bmatrix} = \begin{bmatrix} l_2 \\ \dot{m}_2 \end{bmatrix} - d^{(\nu-1)} \begin{bmatrix} n_2 & \sum_2 x \\ \sum_2 x & \sum_2 x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_1 w\left(\dfrac{u^*-l_2^{(\nu-1)}-m_2^{(\nu-1)}x}{d^{(\nu-1)}}\right) \\ \sum_1 xw\left(\dfrac{u^*-l_2^{(\nu-1)}-m_2^{(\nu-1)}x}{d^{(\nu-1)}}\right) \end{bmatrix}, \tag{6.11}$$

$$\begin{aligned} n(d^{(\nu)})^2 = {} & \sum_1(u^*)^2-l_1^{(\nu)}\sum_1 u^*-m_1^{(\nu)}\sum_1 xu^* \\ & -d^{(\nu-1)}\sum_1 u^* w\left(\frac{u^*-l_2^{(\nu-1)}-m_2^{(\nu-1)}x}{d^{(\nu-1)}}\right) \\ & +\sum_2(u^*)^2-l_2^{(\nu)}\sum_2 u^*-m_2^{(\nu)}\sum_2 xu^* \\ & -d^{(\nu-1)}\sum_2 u^* w\left(\frac{u^*-l_1^{(\nu-1)}-m_2^{(\nu-1)}x}{d^{(\nu-1)}}\right). \end{aligned} \tag{6.12}$$

The final equation also has an analogy with ordinary least-squares estimation. The two terms involving the w function on the right-hand side may be interpreted as adjustments to the other six terms, which constitute the pooled residual sums of squares associated with the fitting of ordinary regression lines to the two scatter diagrams.

Table 2 shows the results of iterating to the maximum-likelihood estimates from initial estimates set equal to the least-squares estimates $l_1, \dot{m}_1, l_2, \dot{m}_1, d$. A similar iteration from the true values of the parameters also took five cycles to reach the

same accuracy. Note that there is an appreciable difference between the maximum-likelihood estimate of μ_1 and the unadjusted least-squares estimate. The slope of the line on which the latter is based is lowered by the admission of the ghost data.

TABLE 2

Iteration to the maximum likelihood estimates

ν	$l_1^{(\nu)}$	$m_1^{(\nu)}$	$l_2^{(\nu)}$	$m_2^{(\nu)}$	$d^{(\nu)}$
0	6·56	0·482	3·38	0·782	0·402
1	7·21	0·361	3·10	0·796	0·408
2	6·70	0·453	3·20	0·791	0·402
3	7·02	0·396	3·18	0·793	0·405
4	6·87	0·422	3·18	0·792	0·404
5	6·88	0·421	3·18	0·792	0·404

6.6. *Optimum Treatment Expectation Model*

In this model the decision-maker records only the number of the treatment which he regards as optimum. Thus he will allocate an individual in initial state x to treatment 1 if and only if the utility estimate u_1 is greater than the utility estimate u_2. Since u_1 and u_2 are in our underlying model independent $N(\lambda_1+\mu_1 x, \delta^2)$ and $N(\lambda_2+\mu_2 x, \delta^2)$ random variables this means that the probability of recording treatment 1 is

$$\Pr\{u_1 > u_2\} = \Phi\left(\frac{\lambda_1-\lambda_2+(\mu_1-\mu_2)x}{\delta\sqrt{2}}\right)$$
$$= \Phi(A+Bx), \tag{6.13}$$

where

$$A = \frac{\lambda_1-\lambda_2}{\delta\sqrt{2}}, \quad B = \frac{\mu_1-\mu_2}{\delta\sqrt{2}}. \tag{6.14}$$

Thus the decision-maker sets

$$\tau(x) = \begin{cases} 1 & \text{with probability } \Phi(A+Bx), \\ 2 & \text{with probability } 1-\Phi(A+Bx), \end{cases} \tag{6.15}$$

and so is operating a randomized allocator as defined in Section 4.3.

This is precisely the basic binomial trial model of *probit analysis* (Finney, 1947) with the choice of treatment 1 as "response" and of treatment 2 as "non-response" to "stimulus strength" x. We have thus a familiar estimation problem with, however, one difference. The drastic reduction in the available information has resulted in the unidentifiability of the basic parameters, for there are clearly many sets of $(\lambda_1, \mu_1, \lambda_2, \mu_2, \delta)$ which give the same (A, B) and hence the same model. Indeed only the forms A and B and functions of them are identifiable. We shall see, however, that the estimation of A and B is sufficient for one important purpose.

The corresponding model for more than two treatments (Aitchison and Bennett, 1970) is more complex and the problem of identifiability is non-trivial and of estimation non-standard.

The true values of A and B in our simulated example are

$$A = 7{\cdot}07, \quad B = -0{\cdot}707.$$

The standard iterative estimation procedure of probit analysis applied to the data give maximum-likelihood estimates

$$\hat{A} = 7{\cdot}11, \quad \hat{B} = -0{\cdot}726.$$

6.7. *Discussion of the Three Models*

Because of the unidentifiability, and hence the inestimability, of the basic parameters of the model of Section 6.6, we cannot compare the standard errors of the estimators to obtain measures of our ability to recover the utility function from the diminishing quality of the information associated with the three models. There is, however, a function of the parameters of particular interest. A decision-maker who could state his utility structure explicitly would choose treatment 1 if and only if $\lambda_1+\mu_1 x > \lambda_2+\mu_2 x$ and so the initial state

$$-\frac{\lambda_1-\lambda_2}{\mu_1-\mu_2} = 10$$

provides a *critical point*. For all x below it one treatment (in our example, 1) is optimum; for all x above it the other treatment (2) is optimum. This critical point, being associated with the identifiable function $-B/A$, is estimable from the data of the probit model. Indeed in the standard terminology of probit analysis the critical point is the ED 50.

Table 3 gives the estimated critical points obtained through the three models together with their estimated standard errors obtained by standard maximum-likelihood approximation methods. The increasing magnitude of the standard errors measures the decreasing effectiveness of the data.

TABLE 3

Estimated critical points and their estimated standard errors

Expectation model	*Estimated critical point*	*Estimated standard error*
Full utility (Section 6.4)	9·74	0·33
Optimum utility (Section 6.5)	9·97	0·43
Optimum treatment (Section 6.6)	9·79	0·60

Even with probit-type data there is clearly some hope of effective estimation, at least of this important critical point. The problem in this case has a deceptively familiar form. For consider the graphical representation of such data for a two-dimensional initial state (x_1, x_2). In Fig. 2 each point represents the initial state of an individual presented to the decision-maker and shows the chosen treatment. For the normal linear utility model the critical point is now replaced by a critical line but again the appropriate estimation procedure is a form of probit analysis. The diagram, however, has the appearance of data for the standard classification problem, the classes being treatments, and it would be tempting to apply discriminant analysis, an entirely different procedure, to the two clusters to determine an appropriate dividing line. If,

in a medical context, there were a one-to-one correspondence between disease and treatment, it might be difficult for the physician to disentangle the classification (diagnosis) and treatment allocation problems. It would be interesting in such a situation to compare the two techniques of discriminant analysis and utility estimation to see what practical differences might arise and possibly to relate misclassification probabilities with utility structures. Such an exercise is beyond the scope of this paper.

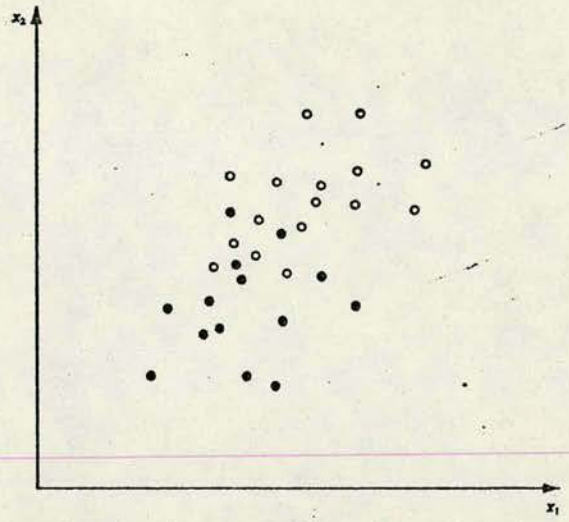

FIG. 2. Initial states (x_1, x_2) and treatments allocated by an inconsistent decision-maker. Treatment 1, ●; treatment 2, ○.

Although we have discussed three different models these relate to differences in what the decision-maker is able to reveal and not in his supposed process of decision-making. The measure of inconsistency of this process or its suboptimality is thus the same for all three models. Suppose that the natural distribution of initial states is normal with mean 11 and standard deviation 5. The suboptimality of our simulated decision-maker is then, by (4.10),

$$\begin{aligned} &I(\lambda_1-\lambda_2, \mu_1-\mu_2, \lambda_1-\lambda_2, \mu_1-\mu_2)-J(\lambda_1-\lambda_2, \mu_1-\mu_2, \lambda_1-\lambda_2, \mu_1-\mu_2) \\ &\quad = \delta\sqrt{2}\{I(A,B,A,B)-J(A,B,A,B)\} \quad \text{with} \quad \psi = 11, \Omega = 25 \\ &\quad = 0{\cdot}030. \end{aligned} \tag{6.16}$$

The estimated suboptimality obtained by the substitution of estimates, say from the optimum utility data, for parameters is 0·032. The average utility per allocation with the optimum allocator

$$\tau^*(x) = \begin{cases} 1 & (x \leqslant 10), \\ 2 & (x > 10), \end{cases}$$

for such a stream of initial states is, by (4.4),

$$\lambda_2 + 11\mu_2 + \delta\sqrt{2I(A,B,A,B)} = 12{\cdot}89,$$

a standard for comparison.

Whether or not a suboptimality of this magnitude is of practical significance must, of course, depend on the circumstances of the real problem. The suboptimality depends, as we would expect, on δ, the error standard deviation, and in our example the magnitude of δ is moderate. Notice also that (6.16) cannot be written solely in terms of A and B, and hence the suboptimality, regarded as a parametric function, is unidentifiable and in consequence inestimable in the optimum treatment expectation model. One interpretation of this is in terms of Fig. 2. If there is considerable overlap of the treatment "clusters" we cannot say to what extent this arises from great inconsistency (large δ) or from the inherent difficulty of decision-making with close mean utility functions $\lambda_1+\mu_1 x$ and $\lambda_2+\mu_2 x$.

It is interesting to quote here the suboptimality of a decision-maker who claims that his utility structure is the same as that of our simulated decision-maker but who insists in allocating only one of the treatments, say 2, for all initial states. This is the constant allocator of Section 4.4 with suboptimality

$$\delta\sqrt{2I(A, B, A, B)} = 1{\cdot}09,$$

compared with the 0·030 of the simulated decision-maker.

6.8. *Prediction Models*

The preceding sections have discussed models for a decision-maker who estimates the expected utilities associated with the separate application of the two treatments. It is also possible to construct models in which the decision-maker is visualized as predicting the final states and hence the actual utilities associated with these applications. For given x and t the prognosis distribution $p(y|x, t)$ induces a distribution for the statistic $U(x, t, y, \zeta)$. For example, with the univariate normal linear utility model this distribution is normal with mean $U(x, t, \zeta)$ or $\lambda_t+\mu_t x$ and standard deviation $|\eta|\sigma$. With such a prediction model allocation to treatment will then be made on the basis of the larger of the utilities predicted. The only difference between prediction and expectation models is therefore that where we previously had δ we now have $|\eta|\sigma$. From the point of view of estimation, the utility structure that we recover will be identical for the two types of model. If we estimate on the assumption that estimated expected utilities are being quoted or used we obtain exactly the same maximum-likelihood estimates of $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ or $(\xi, \eta, \kappa_1, \kappa_2)$ as we would obtain using the corresponding prediction model. The maximum-likelihood estimates $\hat{\eta}$, d and s of η, δ and σ would satisfy the relationship $d = |\hat{\eta}|\, s$. Thus while the two types of model differ in concept there are no practical differences between them.

Another source of inconsistency could be the existence of errors in, or uncertainty about, the initial state vector x. While I have not studied this in any detail I suspect that it will again lead back to something very like the expectation models of earlier sections.

7. Discussion

It must be re-emphasized that this paper is essentially a feasibility study. The main investigation is of the circumstances of treatment allocation under which the utility structure of a decision-maker may be satisfactorily estimated from his behaviour pattern. The work was motivated by proposals that such behavioural studies of certain areas of medical practice might be undertaken. In medicine growing attention to the construction of quantitatively based prognosis distribution raises the question of whether the natural hesitancy of medical decision-makers to make explicit their utility structures can be compensated.

Our illustrative example concerned a one-dimensional initial state and a not-too-inconsistent decision-maker. For higher-dimensional initial states and more inherent inconsistency the amount of data required for satisfactory estimation will naturally increase. Fortunately the greatest need for knowledge of the utility structure is likely to occur in those situations where allocations have to be made repeatedly and where, therefore, more data should be available. Again we have investigated in depth only two-treatment normal linear models. As indicated earlier, the case of more than two treatments raises substantial, though not insuperable, new problems, and the extension to categorical initial and final states is also of interest.

To automate optimum treatment allocation as defined in this paper the estimation of critical points or critical lines is all that is necessary. The hope that at least this estimation will be of practical value is encouraged by the analogy we drew with the problem of classification, and its increasing success in applications to medical diagnosis.

I am grateful to Professor W. I. Card and Dr T. R. Taylor, of the University of Glasgow Department of Medicine in Relation to Mathematics and Computing, for discussion of the medical problems which motivated this paper.

REFERENCES

AITCHISON, J. (1970a). *Choice against Chance*. Reading, Massachusetts: Addison-Wesley.

—— (1970b). Decision making in clinical medicine. *J. R. Coll. Phys., Lond.*, **4**, 195–202.

AITCHISON, J. and BENNETT, J. A. (1970). Polychotomous quantal response by maximum indicant (to appear).

AITCHISON, J. and SCULTHORPE, DIANE (1964). Some problems of statistical prediction. *Biometrika*, **52**, 469–483.

BENDER, I. B., SELTZER, S. and SOLTANOFF, W. (1966). Endodontic success—a reappraisal of criteria. *Oral Surg., Oral Med. and Oral Path.*, **22**, 780–802.

BOYLE, J. A., GREIG, W. R., FRANKLIN, D. A., MARDEN, R. McG., BUCHANAN, W. W. and McGIRR, E. M. (1966). Construction of a model for computer-assisted diagnosis: application to the problem of non-toxic goitre. *Quart. J. Med.*, **35**, 565–588.

CARD, W. I. (1967). Towards a calculus of medicine. *Medical Annual*, pp. 9–21.

—— (1970). The diagnostic process. *J. R. Coll. Phys., Lond.*, **4**, 183–187.

DE GROOT, M. H. (1968). Some problems of optimal stopping. *J. R. Statist. Soc.* B, **30**, 108–132.

FINNEY, D. J. (1947). *Probit Analysis*. Cambridge: University Press.

GINSBERG, A. S. and OFFENSEND, F. L. (1968). An application of decision theory to a medical diagnosis–treatment problem. *I.E.E.E. Transactions on Systems Science and Cybernetics*, SSC **4**, 355–362.

HILLS, M. (1966). Allocation rules and their error rates. *J. R. Statist. Soc.* B, **28**, 1–31.

HUGHES, W. L., KALBFLEISH, J. M., BRANDT, E. N. and COSTILOE, J. P. (1963). Myocardial infarcation prognosis by discriminant analysis. *Archives of Internal Medicine*, **111**, 338–345.

KENNEDY, G. D. C. and SIMPSON, M. S. (1969). The hollow tube controversy. *J. Brit. Endod. Soc.*, **3**, 25–30.

LEDLEY, R. S. and LUSTED, L. B. (1962). Medical diagnosis and modern decision making. In *Mathematical Problems in the Biological Sciences*. Proceedings of Symposia in Applied Mathematics, **14**, 117–158.

LUSTED, L. B. (1968). *Introduction to Medical Decision Making*. Springfield, Illinois: Thomas.

MARSHALL, A. W. and OLKIN, I. (1968). A general approach to some screening and classification problems. *J. R. Statist. Soc.* B, 30, 407–443.

NORRIS, R. M., BRANDT, P. W. T., CAUGHEY, D. E., LEE, A. J. and SCOTT, P. J. (1969). A new coronary prognostic index. *Lancet*, **1**, 274–281.

PEEL, A. A. F., SEMPLE, T., WANG, I., LANCASTER, W. M. and DALL, J. L. C. (1962). *Brit. Heart J.*, **24**, 745–760.

RAIFFA, H. (1968). *Decision Analysis*. Reading, Massachusetts: Addison-Wesley.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University.

SELTZER, S. and BENDER, I. B. (1965). Cognitive dissonance in endodontics. *Oral Surg., Oral Med. and Oral Path.*, 20, 550–516.

SELTZER, S., BENDER, I. B., SMITH, J., FREEDMAN, I. and NAZIMOV, H. (1967). Endodontic failures —an analysis based on clinical, roentgenographic and histologic findings. *Oral Surg., Oral Med. and Oral Path.*, 23, 500–530.

STORMS, J. L. (1969). Factors that influence the success of endodontic treatment. *J. Canad. Dent. Ass.*, 35, 83–97.

TAYLOR, T. R. (1970). Computer-guided diagnosis. *J. R. Coll. Phys., Lond.*, 4, 188–195.

THOMSON, J. D. (1968). *Electricity in Hospitals.* University of Glasgow Building Services Research Unit: Technical Paper No. 7.

APPENDIX

The following integrals make repeated appearances in computations of sub-optimality associated with normal linear models.

Integral 1

Let the n-dimensional random vector x have an $N(\psi, \Omega)$ density function denoted by $\phi(x|\mu, \Omega)$. Then

$$I(a,b,c,d) = \int_{c+d'x\geqslant 0} (a+b'x)\,\phi(x|\psi,\Omega)\,dx$$

$$= (a+b'\psi)\,\Phi\left\{\frac{c+d'\psi}{(d'\Omega d)^{\frac{1}{2}}}\right\} + \frac{b'\Omega d}{(d'\Omega d)^{\frac{1}{2}}}\,\phi\left\{\frac{c+d'\psi}{(d'\Omega d)^{\frac{1}{2}}}\right\}, \qquad \text{(A1)}$$

where ϕ and Φ without displayed parameters denote the standardized univariate normal $N(0, 1)$ density and distribution functions. Moreover,

$$\max_{c,d} I(a,b,c,d) = I(a,b,a,b).$$

This result is readily established by standard techniques, first by a transformation to the $N(0, I_n)$ case, then by the introduction of a suitable orthogonal transformation and finally by standard differential calculus maximization techniques.

A second useful integral which succumbs to the same technique is the following.

Integral 2

$$J(a,b,c,d) = \int_{\boldsymbol{X}} (a+b'x)\,\phi(c+d'x)\,\phi(x|\psi,\Omega)\,dx$$

$$= (a+b'\psi)\,\Phi\left\{\frac{c+d'\psi}{(1+d'\Omega d)^{\frac{1}{2}}}\right\} + \frac{b'\Omega d}{(1+d'\Omega d)^{\frac{1}{2}}}\,\phi\left\{\frac{c+d'\psi}{(1+d'\Omega d)^{\frac{1}{2}}}\right\}. \qquad \text{(A2)}$$

DISCUSSION ON PROFESSOR AITCHISON'S PAPER

Dr M. HILLS (London School of Hygiene and Tropical Medicine): Professor Aitchison's paper raises a number of questions, most of which are concerned with the potential usefulness of his approach in the field of medicine. Before considering any of them I should like to make the general point that there is a considerable danger of distortion when parts of the logic of medicine are formalized in order to make explicit their probabilistic basis. For example, Professor Aitchison, in describing diagnosis as a process whereby abnormalities in a symptom vector are said to be due to the presence of some disease, is distorting the process of diagnosis. Although the paper is about treatment allocation rather than diagnosis, the two are so closely connected that it is very difficult to discuss the one and ignore the other.

It seems to be generally accepted that the closest one can get to a definition of disease is a consensus of opinion about which of a huge number of possible vectors, x, referring to a person's physical and mental state, shall be described as diseased. This set of vectors is then classified in various ways, the most important being by (a) aetiology, (b) functional pathology and morbid anatomy and (c) clinical description. Diagnosis is the process of classifying a particular vector x in one or more of those schemes, and the most common decision which needs to be made in this process is whether or not to make a further investigation of a patient in order to improve the classification. The reason for trying to improve the classification is to narrow the range of possible treatments and to increase the probabilities of favourable outcome to treatment. The diagnostic process may well stop at a relatively crude level if a single treatment is indicated or if the prognosis of a number of treatments is approximately the same. Now this is all relevant to the paper this evening because we are asked to consider a situation in which there are a number of possible treatments, a range of x vectors, and a decision has to be taken about which treatment goes to which x. It is natural to ask when such a situation might arise and one possibility is when the diagnostic procedure has reached the end of the line without narrowing the choice of treatments to one. Another is where the diagnosis stops at a fairly crude level due to ignorance (as in psychiatry). In most situations falling into these two categories the prognosis distribution is not likely to be very informative, although the utilities may vary considerably, and I am led to the conclusion that it is this combination which really interests Professor Aitchison. Taking his example 3(iii), the controversy implies that $p(y \mid x, t)$ is not very informative, but the utility of a treatment which saves a tooth might be considerable to an actor yet low for a hermit. It is very reasonable to assume that in the absence of much information from $p(y \mid x, t)$ the utilities must play a part (albeit unconscious) in the treatment allocations.

This brings us to the real point of the paper, which is the estimation of utilities.

It is suggested that this might serve two purposes, which I shall refer to as confrontation and replacement. If all that is known about a doctor is his treatment allocation, $t^* = t_1$ or t_2, for a range of x vectors, then the technique of discriminant analysis (or probit analysis or logistic regression) expresses $P(t^* = t_1 \mid x)$ as some simple function of a linear form in x. Such an empirical relationship might be used to show that the allocation behaviour was strongly influenced by social class and ages and not much else, and this might be a surprising "confrontation" for the doctor. If the fit of the model was very good it could be used for future allocation, perhaps even replacing the doctor, although this would obviously require extensive data over the range of possible x's. I should like to separate the application from the methodology here because the confrontation seems to me to be most interesting and potentially valuable whereas the use of parametric utilities does not seem to add to the standard statistical methods for this problem. The other models, in which the doctor is able to state utilities, I find difficult to visualize. The initial step of stating utilities at all seems to me to be so large that the question of whether or not one smooths them by regression on x pales into insignificance.

In conclusion I would suggest that Professor Aitchison is dealing with a much more restricted situation than he would have us believe from his very general formulation, but that in this situation there is considerable interest in the further analysis of allocation behaviour. An analogy may be drawn with voting behaviour which has, incidentally, been intensively analysed without bringing in utility functions. That I am unable to follow Professor Aitchison the whole way with his utility models is probably due to a personal view that such models tend to impose a rigidity which is not compatible with the changing state of medical knowledge and availability of new treatments. The news that statisticians are actively engaged in considering these problems may, I think, be received with mixed feelings by the medical profession, but this evening our own feelings can only be those of pleasure in the interesting possibilities which Professor Aitchison has revealed to us, and I am happy to propose a vote of thanks.

Professor D. R. Cox (Imperial College, London): The present status of statistical decision theory is rather puzzling. If utility can be identified with money and all the probability distributions are effectively based on data, the position is clear. Important examples, for instance in control theory, acceptance sampling and in stock control certainly exist, but these are a small proportion of practical decision problems. An elegant and apparently all-embracing general theory is available involving subjective probability and subjective assessments of utility. Yet I, for one, am unconvinced that this necessarily captures the essence of decision problems. (For instance, does Example 5 of the paper do more than reword the Chancellor of the Exchequer's problem?) Professor Aitchison's paper is doubly welcome for the interesting new technical material and for the balanced discussion of what are clearly a group of challenging and important practical problems.

I want to make two detailed comments and then to return to the general question of formulation.

There is a connection with familiar problems in experimental design. If the treatments are indexed by one or more continuous variables, an allocator is a choice of optimum conditions on the basis of data. Yates (1952) and subsequently Anderson and Dillon (1968) have investigated the economic and decision-theoretic aspects of this, but further investigation from the viewpoint of Professor Aitchison's work might be fruitful.

Next, in the interesting estimation problem connected with Fig. 1 the use of a logistic density rather than a normal would have simplified the likelihood and combined with the approximation $e^x/(1+e^x) \doteq 1+kx\,(|x|<1/k)$ might have led to fairly simple estimates from which to begin an iteration; this would be especially worth considering in situations more complicated than that of (6.8). A further statistical problem associated with this sort of data concerns an appropriate definition of residuals for testing goodness of fit.

Returning now to the general issues, I think that Professor Aitchison's explicit introduction of a prognosis distribution is an important clarifying idea. In his Section 5 Professor Aitchison is, in more familiar terminology, suggesting the estimation of treatment differences and of their interactions with important classificatory variables; the emphasis on interactions is important. Professor Aitchison has, however, put main emphasis on the utility function. There are, in essence, two extreme forms of decision problem. One is typified by the view of life in which one is constantly faced with clear-cut conflicts between selfish and altruistic considerations, i.e. the choice of utility is at stake. On the other hand, the uncertainty may be entirely in assessing the consequences of any decision taken, i.e. the choice of prognosis distribution may be central. Where in between these extremes do the practical problems considered by Professor Aitchison lie? Even if the difficulties lie primarily in the choice of utility, should not much more emphasis be placed on establishing the components of utility? What are the main components; how do they depend on the final state y; what relative importance is attached to these components first by patients and then by doctors; is a rough completely "economic" weighting feasible, etc.?

Quite generally, I feel that much of statistical decision theory can be criticized for its concentration on a premature *synthesis* of utility and subjective probability into single numbers, when the emphasis should be on the *analysis* of these concepts into components and the detailed study of these. Synthesis there must be when the final decision is to be made, but the gaining of understanding and the improvement of decision-making seem to require much more than an immediate sweeping together of these difficult concepts into single numbers.

I want to attach low utility to the manufacture of sources of disagreement that do not really exist and there is nothing in Professor Aitchison's paper, or indeed in the literature on decision theory, to suggest that it would not be a good idea to make the above kind of analysis. The point is that this does not get much emphasis.

Professor Aitchison has given a lucid and interesting paper on an important topic and I have much pleasure in seconding the vote of thanks.

The vote of thanks was put to the meeting and carried unanimously.

Dr J. A. ANDERSON (Department of Biomathematics, Oxford University): I would like to say how much I have enjoyed Professor Aitchison's paper, particularly since I found it very thought-provoking on several counts. He is to be congratulated on a clear and thorough presentation of the structure of the treatment allocation problem. However, I would like to question one or two points he made. His chief objective is to estimate satisfactorily the utility structure in treatment allocation. Presumably, his method would be judged to be successful if it predicted accurately decisions actually made by a doctor. This is clearly a very interesting field of research but it suffers from some drawbacks. Firstly, if the doctor's and the system's decisions disagree, it could be due to inadequacies in either but how will we know which? Secondly, it seems a little restricting to try to reproduce one doctor's decisions. Why not allow the system and the doctor to improve, if possible? This would suppose that there exist objective criteria for judging the success of a treatment. In medicine, where decisions are being made repeatedly, I think this is quite possible.

I was amused to see quality control and management of the economy included as examples of medical situations suitable for treatment allocation. Irrespective of how reassuring we find the parallel between doctors choosing treatments and politicians choosing economic strategies, it seems that Professor Aitchison regards them as equivalent for decision-making purposes. I think there are two differences; firstly, medical decisions are made repeatedly whereas I believe that economic strategies are varied infrequently, and secondly, the pay-off from an economic decision is measured almost entirely in financial terms, while in medicine at least three distinct elements can be distinguished:

(1) the patient's survival,
(2) the quality of the patient's remaining life,
(3) the financial cost.

Although there are strong theoretical reasons why a decision-making system should depend on a single utility function, as in the present paper, there are practical reasons in the medical context for overriding this requirement.

I think most doctors would find a utility function, in arbitrary units, rather artificial, so I suggest that an acceptable treatment allocation system might be to present the risks (factors 1 and 2) and the financial cost associated with each treatment, leaving the doctor to make the final decision. A doctor's decision is based on considerations like my factors, so why not give him objective assessments of them? This might be the first stage, the second stage being to use the factors in a statistical system. This could be as suggested here, or based on a linear combination of my factors (Fishburn, 1964) or even based on putting bounds on the various risks as suggested by myself (Anderson, 1969).

There are several reasons why I think rules based on the factors (1)–(3) are to be preferred. Firstly, they are quite objective quantities, so the associated risks can be estimated without any further assumptions. In fact, this exercise would be very similar to estimating Professor Aitchison's prognosis distributions. Although I have no hope of sample sizes large enough to ignore sampling errors, at least something can be said about them. I like the device of assuming that guesses of utilities have a given error distribution, particularly since it gives such satisfying results with the normal linear model. However, I cannot help wondering about the size of the sampling error in the resulting utility estimates, bearing in mind that in real-life the unknown prognosis distribution also has to be estimated.

It may be that the procedure of the doctor deciding between statistically assessed risks might be more advantageous in the long-term, since it is more flexible. For example, a physician choosing between surgery and radio-therapy might be influenced by the availability of a particular surgeon and make adjustments to the appropriate risks. In another context, perhaps kidney disease, a deciding factor might be the patient's willingness or

ability to find a large sum of money. In this case it might be appropriate for the patient to be told the risks and for him to make the decision.

In these circumstances, I suggest that the objective and subjective stages in decision making should be kept separate.

Another point to remember is that several doctors might be concerned in the treatment of a patient, making one set of personal utilities unworkable.

Professor W. CARD (University of Glasgow): I am grateful for the opportunity of commenting on this very interesting paper. I am a doctor and not a statistician so I obviously cannot comment on the technical mathematics but only on those parts which particularly affect clinical medicine.

In his formulation Professor Aitchison speaks of a final state y and a utility function of x, t and y. This rather suggests that, in terms of medicine, he has in mind some acute condition which can be said to have a final state. But we are frequently concerned with treating a patient where the outcome cannot ever be regarded as final but has a duration measured in years, e.g. freedom from recurrence of a cancer. It seems to me therefore that the utility state has to be considered as having a duration over a period of time, and that this in itself has some further probability distribution. Doctors get over this by putting the patient into a particular class. Medicine perhaps is unusual in that the subjective awareness of the state y is only part of the total utility state of the subject.

Professor Aitchison classifies three types of utility expectation models, but it is not easy to believe that any doctor at the moment would be prepared to estimate utilities numerically as demanded by his first and second models. There is no doubt that the doctor could name what he believed to be his optimum treatment as in Model 3. But I think a doctor could go further and could express his preferences over a series of utility states and give some estimate of his degrees of preference (Card and Good, 1970). Though we find it difficult to think in terms of numerical estimates of utilities today we may learn to do so in the future and here one might suggest an analogy with the percentage disability, decided by a Medical Board under the Industrial Injuries Act. In effect this is a numerical estimate of a utility. I also think it is important to point out that we should get a good deal of help if the methods Professor Aitchison proposes only allowed us to calculate big differences in expected utility as this would give us more time to spend on the subtle decisions.

Professor Aitchison has made the simplifying assumption that his two treatments, t_1 and t_2, are mutually exclusive. But often in medical practice this is not so. If we have an alternative medical and surgical treatment as for example in duodenal ulcer, we may advise medical treatment first in the knowledge that surgical treatment can always be carried out at a later date. Also if we have two medical treatments it is not uncommon for them to be carried out sequentially especially when a treatment may itself provide a valuable piece of diagnostic evidence, the so-called therapeutic test of the diagnosis.

Finally, may I say how gratifying it is that statisticians like Professor Aitchison are prepared to interest themselves in some of our medical problems which involve decisions.

Mr M. J. R. HEALY (Medical Research Council): I would like to add my congratulations to Professor Aitchison for his lucid and interesting paper. The neglect of the problem of treatment allocation is certainly remarkable; it is becoming increasingly irksome in the field of clinical trials, where the statistician's instinct to relegate treatment × subject interaction in its entirety to error (however reasonable this may be in agricultural field trials) runs counter to the clinical fact that different categories of patients respond differently to almost any particular treatment.

Where I would differ to some extent from Professor Aitchison is in the relative emphasis which he gives to the three components of the problem—the initial state, the

prognosis distribution and the utilities. It is of interest that a practitioner's implicit system of utilities can to some extent be deduced from the practice—it is arguable that a rational man ought to use some system of utilities, and it could be useful to demonstrate inconsistencies. However, I am not sure in what sense a utility function can be said to be *wrong*; on being made aware of my implicit utilities I may wish to change them, but I feel that irreducible differences of opinion are likely to persist as between, for example, different doctors or between doctor and patient.

By contrast, the prognosis distribution is in essence objective and subject to investigation, as Professor Aitchison points out. Actually obtaining information may, however, be exceedingly difficult in areas such as medicine in which the possibilities of experiment are limited—the outlook for getting the prognosis distribution from non-experimental observations is not bright, as has been clearly shown in agricultural fertilizer work where the role of the prognosis distribution (or at least its expectation, given t) is played by the response curve. Equations (5.1) and (5.2) in the paper are a little misleading without some indication that the left-hand sides are estimates subject to more or less uncertainty. Dr Yates's *Nature* article, to which Professor Cox referred, was essentially concerned with the costs involved in estimating the prognosis distribution, and it is of interest that he pointed out the value of studying its dependence on the initial conditions so as to avoid the type of suboptimality described in Section 4.5.

The third part of any particular allocation problem is the initial state. This differs from the prognosis distribution in that it must be determined afresh for each new case and consequently the costs associated with its determination cannot be avoided in the overall utility reckoning. Fertilizer work again provides an example—much argument has been expended on the accuracy to which soil analyses should be determined, but this is largely misplaced in the absence of knowledge of the full prognosis distribution which would render the utilities (here fairly simply arrived at) calculable. In the medical context, the cost (to both the patient and the community) of elaborate investigations can be large, and decisions about them may be as important as treatment decisions.

Dr H. THOMAS (London Graduate School of Business Studies): I would like to add my thanks to Professor Aitchison for his paper. I do not want to discuss any aspect of the medical problem because I know nothing about it. I have two general points—one is about the practical applications of the type of formulation that Professor Aitchison has put forward, and the other concerns one possible practical application in the business studies context.

My first point ties in with what Professor Cox was saying. This centres round the confusion existing in statistical decision theory and was summed up by Professor Ehrenberg at the Sheffield R.S.S. conference last year. It was pointed out there that since there appear to be few useful routine applications of Bayesian procedures in the literature, there seems to be some doubt about whether Bayesian procedures are applicable at all to practical problems. I do not object to Professor Aitchison's formulation at all because I think it presents the decision-theory framework in a different way, and in an interesting way, but it adds to the rather sterile debate on a theoretical level about statistical decision theory and does not furnish us with any larger applications of Bayesian procedures in practice.

My second point is on its possible application to the type of investment decision problem which Professor Plackett was considering in a paper he read to this Society recently. I have been doing some work on this particular problem, trying to estimate such things as the prognosis distribution and the utility function given that you have some initial state of an investment project (e.g. initial estimated rate of return) which is x and have some treatment factor t (men, money, machines) and you want to arrive at some final state, a final rate of return or something of this nature desirable to a business firm. In this

context, we meet again the problem of how to obtain information on the prognosis distribution; many business men will tell you that every investment decision is different from every other, so one has no historical data base and, therefore, no objective evidence to enable one to assess a prognosis distribution.

The other problem that occurs in the application of Professor Aitchison's structure to an investment decision problem is the relevant utility structure for the problem: it seems to me that if one's objective, as Professor Aitchison has pointed out, is to attempt in some way to automate the decision-making process, why should we be trying to recover the information from how the individual decision maker makes the decision in a given context? The effort should be placed on obtaining a corporate global utility function. I apologize for raising this application in a basically medical discussion, but I feel it is an important issue: it would be nice if we could have some routine practical applications of Bayesian methods in the *Journal* as well as the debate on statistical decision theory. One application I should commend to attention in the business field is the thesis of Gittins on optimal resource allocation written in 1968. I believe Professor Davies is doing work in the chemical industry on this problem.

I thank the author for the paper. The structure is very useful for people working in the practical area.

Dr T. R. TAYLOR (University of Glasgow): I would like to contribute to this discussion as a physician who is actively interested in the application of decision theory to clinical problems and who is collaborating with Professor Aitchison in this field in Glasgow.

As I see it, the aims of those working in this field should be threefold:

(1) To develop a detailed model of the diagnostic process including treatment allocation.
(2) To pursue the analysis of decisions made by physicians in the course of their day-to-day work so that some of the simpler ones may be automated and insight gained into the more complex ones.
(3) Attempt to extend our analysis of the decisions taken by physicians to allow them to learn more about their own decision-making behaviour: this is referred to by Professor Aitchison as "confrontation".

We will depend, for the success of these aims, on convincing our fellow physicians of the practical as well as theoretical value of decision theory. We must therefore choose an operational decision unit which is clinically meaningful as well as being tractable from the design point of view to the statistician. For these reasons we have chosen not an individual decision-maker but a single out-patient clinic.

We have been in fact actively studying the thyroid out-patient clinic in Glasgow Royal Infirmary and have had excellent collaboration from Dr John A. Thomson and his staff and the active encouragement of Professor E. McGirr. In this clinic we have the essential ingredients for a tractable decision system, namely a stream of patients passing through the clinic, a group of decision makers and the ancillary nursing, laboratory and other staff and facilities which are to be deployed in the course of decision making and treatment.

Our approach has been to impose on this unit a structure which will be operationally meaningful to doctors and will fulfil the design criteria necessary for our studies. I myself am a full-time active member of the team in this clinic and I regard this as an essential part of the educational progress without which our decision makers would not be able to collaborate however well disposed they might be. The structure which we have imposed on this decision unit is as follows:

(1) We have produced a set of definitions of all the symptoms, signs, laboratory tests and diseases in the thyroid clinic. These have been amended and have been finally approved by all the physicians involved in the thyroid group at the Royal Infirmary, and our aim is in the near future to attempt to have all the thyroid physicians in Scotland agree on a similar set of definitions.

(2) A prospective survey of all the patients seen in this thyroid clinic is in progress from which we will be able to derive our prior probabilities, likelihoods and the prognosis density function referred to by Professor Aitchison.
(3) A work-study analysis and costing analysis is now in progress to provide us with a financial estimate of the cost of each item of evidence and each treatment used by the decision makers in the clinic. Other "costs" such as the discomfort of the patient and the inconvenience of the investigations are also being investigated by myself by interviewing patients.

As a final comment may I clarify our intentions and answer the criticisms of two of the previous speakers about the final utility structure which we hope to infer from our studies. This utility structure has been referred to as a single item by Professor Aitchison in his theoretical analysis. We are well aware that this is intuitively not the way in which the physician-decision maker thinks about costs. We have analysed costs into (i) financial costs, (ii) the *discomfort* to the patient of his initial state and (iii) a similar analysis of the *inconvenience*, financial or otherwise, of the initial state, the investigations and the treatments and final state. This analysis is at the present moment only theoretical, but we are in the process of pursuing it in some clinical cases in the prospective survey at the present moment.

The value of confronting of the decision-maker with his utility structure is one which will undoubtedly be of value, possibly more so than any other aspect of this work. In a study now nearing completion, we have compared the six physicians in the thyroid clinic in tackling twenty varied cases of non-toxic goitre and the reactions of these physicians to the results of this analysis have been most instructive. Indeed the individual physicians are enthusiastic about this study in improving their performance as diagnosticians and clinical decision-makers.

Professor D. J. NEWELL (University of Newcastle-upon-Tyne): The beginning of this paper is very valuable, because nearly all clinical trials in the past have been designed on the assumption that the population of responses to treatment will be homogeneous, that is that there will be no interaction between the "x-variables" and the response to treatment.

Of the few trials which have considered the possibility of non-homogeneous response, the first that I can recall was by Newton and Tanner (1956). Here the response was preference for one drug or another. Using a double-cross-over design, they devised a statistical analysis (further discussed by Armitage and Healy, 1957) which first finds the most frequently preferred treatment, and then determines whether there is a minority group which consistently prefers the other treatment. It is then open to the investigator to discover whether this group has identifiable characteristics for future treatment allocation.

One of the advantages for Professor Aitchison of trials where the assumption of uniform response has been made is that they sometimes yield data of the sort he is looking for. A trivial example is this: we have recently looked at a particular treatment for the post-infarction period in coronary thrombosis (Newell *et al.*, 1970). Overall (on the assumption of homogeneity) it has no effect on mortality, but when we start looking at the interactions it becomes interesting. Selecting from the x-variables just age and sex, we find that the drug saves the lives of young men, but is apparently lethal to old women. Where I have doubts, in common with previous discussants, is in the allocation of utilities. Using purely accounting utilities, one might say that this is uniformly a good treatment, since it saves the lives of young males who could contribute to national income, and it eliminates some old people whose contribution, in a strictly financial sense, is negative. These are not the sort of utilities that doctors would put on these outcomes. But it is very difficult to quantify utilities in a case of this nature. Similarly, since some doctors have in the past used this treatment for patients of any age and either sex, they would be alarmed if the utilities

implied by their actions were imputed to them. So although the concepts in this paper are admirable, one can foresee great difficulty in the assignment of utilities.

Mr J. G. SKELLAM (Nature Conservancy): Even if a utility function is well defined, there are two aspects of utility—short-term and long-term utility. It may sometimes pay not to aim repeatedly at short-term utility, but to make sub-optimum decisions particularly in the early stages in order to gain knowledge which may throw light, for example, on the prognosis distribution. Extra information and research are invaluable in achieving greater utility afterwards.

The author replied at the meeting and later added to his reply in writing, as follows:

Our Society is probably unique in its approach to treatment allocation. It takes its patient (the reader of a paper), encourages him to disclose his initial state, subjects him almost simultaneously to a number of different treatments, and then awaits the publication of the *Journal* to see him declare his final state. I would like to thank all the discussants for their considerate treatment, which stopped short of the radical surgery I have been conditioned to expect, and for their helpful references to related work. I hasten to assure them that my final state after digesting their prescriptions remains one of uncured optimism.

Before I briefly take up the two basic questions raised in the discussion may I remove a number of the obscurities in my paper which seem to have led to misunderstanding and misinterpretation. First, my purpose in presenting the five illustrative examples was simply to show the relative usefulness of the decision-theory approach in situations of different difficulty. I would thus agree with Professor Cox that in the Chancellor's problem (Example 5) the difficulty is such that formulating it as a decision problem merely reiterates the difficulty; this was the purpose of my choice. But I would certainly not agree with his suggestion that statistical decision theory fails to capture the essence of decision problems. I have seen its practical merits in the tariff selection problem, and already in the thyroid clinic currently being studied we can see the insight which the clinicians are obtaining into their problems. If the theory does nothing else it certainly highlights what information is missing, and amazingly this is one of the features often unrecognized by workers close to a practical situation. I do not understand Dr Anderson's comment about my inclusion of quality control and management of the economy as examples of medical situations suitable for treatment allocation. I did not give them as such examples. The paper is, *in particular*, about the problems of medical treatment allocation, but the issues are broader, as Dr Thomas in his example from research and development indicates.

Although I tried hard in the paper to steer the discussion away from controversy over the merits or demerits of a Bayesian approach Dr Thomas has raised it by expressing doubts about the applicability of Bayesian procedure to practical problems. For the medical applications, which are now my main concern, I would make just one comment. In the studies into decision-making in the thyroid clinic in the Royal Infirmary at Glasgow referred to by Dr Taylor in the discussion we have had no difficulty in persuading clinicians to be Bayesians. They were already Bayesians before the study began, though they might not have called themselves by that name. There is no difficulty in getting them to update their priors after each piece of information becomes available. Moreover, the idea of prognosis, though not its quantification, is in daily use in medical circles. I cannot promise Dr Thomas a large application of Bayesian procedures in medical practice at the moment, but I am convinced that with patience and in the enlightened atmosphere of Professor McGirr's Department in the Royal Infirmary of Glasgow some worthwhile saving of doctor's time and effort is attainable.

My purpose in asking whether one can recover the implicit utility structure actually used is not, as Dr Anderson interprets it, because I think this is the end of the exercise.

It is indeed only the beginning, an attempt to demonstrate that there is something which is implicitly used. A conviction of its existence would perhaps encourage the more direct construction of utility functions. We have seen in Professor Card's discussion that while he cannot envisage any doctor at the moment assessing utilities numerically (although clearly his Medical Board is in fact doing this already) he sees some future hope. On this point, one must ask how far short his degrees of preference are from cardinal utility. Clearly they lie somewhere between ordinal and cardinal, but I suspect that for decision-making under uncertainty he will find that they fall short of adequacy for clear-cut decision-making. Rationality implies the existence of cardinal utility and what he is suggesting falls short of this. It may be a step towards it however.

I am confused by Professor Card's discussion of "final states". At one point he seems to be saying that my formulation cannot deal with situations where the assessment of treatment takes place over time. The term *final* is used to distinguish it from *initial* and the y could denote a realization of a multivariate stochastic process. (I do not claim that the mathematics is easy.) At another point he seems to be claiming that doctors have a way of overcoming this time-involvement by classifying the patients. If doctors are happy with such a classification system then it simplifies my task because it means that Y consists simply of the classes specified by the doctor, and I have a nice finite set to handle. The doctor's mind in fact has grasped the difficulty of the realistic mathematical formulation and has already done the mathematical simplification in converting to classes.

Professor Card's point about mutual exclusion is also due I think to a misunderstanding. It is a familiar device of decision theory that in defining the action set—here the class of treatments—one sets out one's possible actions so that they are mutually exclusive. With Professor Card's two medical treatments, m_1 and m_2 for example, we might consider four or even more new "treatments" to reach this mutual exclusion:

$t_1 : m_1$ only, $t_3 : m_2$ only,

$t_2 : m_1$ followed by m_2, $t_4 : m_2$ followed by m_1.

Of course the more natural way of analysing such a problem would be a formulation as a sequential medical decision problem, reverting to the class $\{m_1, m_2\}$ of mutual exclusive actions. Courses of action can really only be satisfactorily compared if we take steps to make the formulation with this mutually exclusive criterion.

I apologize to Mr Healy for the use of the word "wrong" in connexion with utility functions. "Different" would have been a better and unemotive word. All I was aiming to obtain in Section 4.2 was some measure which would allow me to say to a decision-maker who, when faced with an estimated utility $a+b'x$, said he thought it would be better to use $c+d'x$: "This is what this difference amounts to in terms of expected utilities". Mr Skellam's remark is, I think, interpretable as saying that problems are often sequential. There must indeed be few medical decision problems which are not sequential. What I have been attempting is to obtain an oversimplification which may be tried out in a simple situation. It is the way of all applied mathematics that one builds up from oversimplification towards reality.

The two major points that have been raised are (i) the diagnosis–treatment allocation controversy and (ii) the "all your eggs in one utility basket" or "you cannot put a price on life' syndrome".

(i) It has been interesting to note that the diagnosis–treatment division of medicine has been questioned, not by the two doctors taking part, but by statisticians, in particular by Dr Hills. He seems to concentrate on the semantic problems associated with diagnosis. I agree that I used my language loosely when I said that the cause of abnormalities was the presence of disease. I also agree that the only sensible definition is in terms of the classification of vectors. It follows then that the vector itself or some subset of it or some

extension of it is clearly relevant to the assignment of treatment. Does the term "diagnosis" mean the extension of this vector to the position where classification is synonymous with treatment? Certainly in the practice of medicine today there appear to be many situations where the doctor would suppose that he had completed the diagnosite process and yet there are still a number of available treatments. Of course the search for indicators should continue, and presumably the indicators will further divide the existing disease categories. But it is interesting to note that those engaged in statistical problems of diagnosis seem seldom directly to concern themselves with the continuation of the problem in prognosis and treatment. Have they always been dealing with situations where there is a single clear-cut treatment once diagnosis is completed?

(ii) I am taken to task because I attempt to show to what extent an implicit utility function can be estimated. The mention of a utility function seems anathema to some discussants. Rather than put all your eggs in one basket, study all the components of utility (Professor Cox, Dr Anderson). I think that Dr Taylor has already shown in his description of our work in the thyroid clinic that we are aware of the various components of utility, and I need not enlarge on the details. The intriguing point is how does the decision-maker mould together these various aspects in reaching a final decision. What are the scaling factors used? Now it is possible to sit back and say that this kind of scaling or balancing is best left to the subconscious—the "we must not put a price on life" attitude. What I was attempting to illustrate was how insight into the implicit scaling might be obtained from behavioural studies with various qualities of information. I emphasize again the feasibility aspect of this study. I set as my object the exploration of various qualities of data. The study I think shows clearly what must be demanded to gain what insight. My illustrative example was the simplest I could find to explain how the counter-claims of advantage and cost were scaled in the decision-making process.

It was gratifying to find agreement about the aim of clinical trials and I hope that our discussion tonight may lead to deeper investigations of these.

We are still in the early days of statistical decision theory. It is clear that the ideas of Raiffa, Schlaiffer and their colleagues have had at least a catalytic effect on business and economic decision-making. The utility specification problems there are sizeable and the position is often less promising in their inability to "experiment". In the medical world the $p(y \mid x, t)$ is a recognizable object and it is clear that much more systematic quantification of it is possible. There are arts, for example the basic reference lottery ticket approach, of attempting to elicit numerical utilities, which direct the attention of quite unpromising decision-makers to the right kind of reasoning.

Finally, many of the discussants refer to the great difficulty of the problems. They are also important. We in this country are surely in a unique position with our N.H.S. which is certainly capable of yielding with co-operation and careful forethought, data of immense value for the transformation of unfortunate x's by optimum t's into happier y's.

References in the Discussion

Anderson, J. A. (1969). Constrained discrimination between k populations. *J. R. Statist. Soc. B*, **31**, 123–139.

Anderson, J. R. and Dillon, J. L. (1968). Economic considerations in response research. *Amer. J. Agric. Econ.*, 50, 130–142.

Armitage, P. and Healy, M. J. R. (1957). Query on interpretation of χ^2 tests. *Biometrics*, **13**, 113–115.

Card, W. I. and Good, I. J. (1970). *Mathematical Biosciences* 6, 45–54.

Fishburn, P. C. (1964). *Decision and Value Theory*. New York: Wiley.

Newell, D. J. and clinical collaborators (1970). Penta erythritol tetranitrate (sustained action) in acute myocardial infarction. *Brit. Heart J.*, 32, 16–20.

Newton, D. R. L. and Tanner, J. M. (1956). *N*-acetyl-para-aminophenol as an analgesic. *Brit. Med. J.*, ii, 1096–1099.

Yates, F. (1952). Principles governing the amount of experimentation in developmental work. *Nature*, **170**, 138–140.

AITCHISON, J. and BENNETT, J.A. (1970)

Polychotomous quantal response by maximum indicant

Reprinted from *Biometrika* 57, 253-62

Biometrika (1970), 57, 2, *p.* 253
With 1 *text-figure*
Printed in Great Britain

Polychotomous quantal response by maximum indicant

BY J. AITCHISON AND J. A. BENNETT
University of Glasgow

SUMMARY

Probit analysis can be reformulated in a model where the subject's response or non-response is determined by which one of two random variables, indicants, is the greater. The extension of this formulation to more than two categories of response leads to a new generalization of probit analysis, which raises interesting identifiability and estimation problems.

1. INTRODUCTION

In its simplest form probit analysis (Finney, 1947) is concerned with a basic experiment in which an experimental unit is subjected to a stimulus of known strength and the category of outcome, response or nonresponse, recorded. For a stimulus of strength x the probability that a subject responds is assumed to be

$$\Phi(A+Bx),$$

where Φ is the standardized normal distribution function, and A and B are unknown parameters. The purpose of an informative experiment which subjects n units to a stimulus, the ith unit receiving strength x_i, is simply the estimation of the parameters A and B, or some function of them, for example $-A/B$, the so-called LD 50. The mechanism traditionally visualized in this confrontation between stimulus and subject is that each subject has some specific natural tolerance to the stimulus and will respond if and only if the strength of the stimulus applied exceeds this value. On the assumption that tolerance is $N(\mu, \sigma^2)$, i.e. normally distributed with mean μ and variance σ^2, over the experimental units, the probability that a unit, chosen at random, responds to a stimulus strength x is

$$\Phi\{(x-\mu)/\sigma\} = \Phi(A+Bx),$$

with $A = -\mu/\sigma$ and $B = 1/\sigma$.

There is, however, another way in which this model can be generated and which allows a generalization to more than two categories of response. The generation is particularly appropriate to situations where the category of outcome determined by the experimental unit may be regarded as the result of some reasoning, or possibly subconscious, process of assessing the effects of the choices of the various response categories. An indication of the nature of this kind of model is given by Aitchison (1970) in a situation where the stimulus x is the initial state of some unit and where the two categories are two possible treatments which may be applied to the unit. Here we shall set this initial motivation in terms of a simple economic example.

Suppose that we are studying the demand for some commodity and that the possible responses to the stimulus of an income x are purchasing, response 1, and not purchasing, response 2. We can here imagine the experimental unit, a person, visualizing two separate experiments. The first of these experiments assesses the amount of enjoyment, y_1 say, that

will arise from the purchase of the commodity taking account of the consequential discomfort of being short of money. The second experiment looks forward under the conditions of not having purchased the commodity and assesses the amount of enjoyment, y_2 say, arising from retention of income but lack of the commodity. Suppose that, for a person with income x, the independent random variables y_1 and y_2 are $N(\alpha_1+\beta_1 x, \sigma^2)$ and $N(\alpha_2+\beta_2 x, \sigma^2)$.

A purchase will be made if and only if $y_1 > y_2$. Putting the argument in a form which leads to straightforward generalization, we have, for the probability of a response, purchasing,

$$\begin{aligned}\mathrm{pr}(y_1 > y_2) &= \frac{1}{\sigma}\int_{-\infty}^{\infty} \Phi\{(y_1-\alpha_2-\beta_2 x)/\sigma\}\,\phi\{(y_1-\alpha_1-\beta_1 x)/\sigma\}\,dy_1 \\ &= \Phi[\{\alpha_1-\alpha_2+(\beta_1-\beta_2)x\}/(\sigma\sqrt{2})] \\ &= \Phi(A+Bx). \qquad (1{\cdot}1)\end{aligned}$$

Thus, we arrive at exactly the probit model with

$$A = (\alpha_1-\alpha_2)/(\sigma\sqrt{2}), \quad B = (\beta_1-\beta_2)/(\sigma\sqrt{2}).$$

If we term y_1 and y_2 the indicants of responses 1 and 2, then the choice of the quantal response is by maximum indicant. One point worth noting at this stage is that the mean indicant lines $y = \alpha_1+\beta_1 x$ and $y = \alpha_2+\beta_2 x$ intersect where $x = -(\alpha_1-\alpha_2)/(\beta_1-\beta_2) = -A/B$, that is at the LD 50 strength.

In § 2 we give the generalization of this maximum indicant model to the case of more than two categories of response. This formulation leads to an interesting problem, not normally present in probit analysis, concerning identifiability of parameters. This is already apparent in a simple form in our reformulation of the binary response probit model, for clearly $(\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma)$ is not identifiable although the parametric functions $(\alpha_1-\alpha_2)/\sigma$ and $(\beta_1-\beta_2)/\sigma$ are identifiable. In the more general case the problem is more subtle and its resolution forms the subject of § 3. The process of estimation is discussed in § 4. Our interest in this model arose from the development of techniques for estimating utility functions from decision behaviour, particularly in the allocation of treatments in simple medical situations. Unfortunately the collection of data on such behaviour is a long-term project and so in § 5, to illustrate the estimation procedure, we have had to resort to artificially constructed data. We hope that this will at least demonstrate the feasibility of the approach and perhaps reveal a model of some interest to workers in fields other than medicine.

2. THE GENERAL MODEL

Suppose that there are c categories of response, denoted $1, \ldots, c$. An experimental unit assigned stimulus x performs c independent experiments. The outcome y_i of the ith experiment is the indicant of response category i and is assumed to be distributed as $N(\alpha_i+\beta_i x, \sigma^2)$ $(i = 1, \ldots, c)$. The subject then chooses that category of response which corresponds to the maximum indicant. We assume that there are no ties since the probability of ties is negligible.

It is of course arguable whether the assumption of independence is a realistic one. Since one subject is visualizing the imaginary experiments it would be more reasonable to suppose that the y_i are jointly distributed with nonzero correlations. Our reason for studying the case of independence here is mainly the applied mathematical attitude that when there are tensions between simplicity and realism, and between tractability and intractability, a

simple tractable model may at least be useful in detecting where departures from realism occur whereas a realistic intractable model can serve little purpose. The independent model has sufficient components to explain the inconsistent behaviour of decision makers envisaged by Aitchison (1970) without leading to impossible mathematics. The introduction of correlations would so complicate the inherent identifiability difficulties as to make impossible the considerable estimation difficulties.

The probability that category i is chosen is, therefore,

$$\begin{aligned} p_i(x) &= \mathrm{pr}\{y_i = \max(y_1, \ldots, y_c)\} \\ &= \mathrm{pr}\{y_i > y_j \quad (j \neq i)\} \\ &= \frac{1}{\sigma}\int_{-\infty}^{\infty} \phi\{(y_i-\alpha_i-\beta_i x)/\sigma\} \prod_{j\neq i} \Phi\{(y_i-\alpha_j-\beta_j x)/\sigma\}\, dy_i \\ &= \int_{-\infty}^{\infty} \phi(v) \prod_{j\neq i} \Phi\{v+(\alpha_i-\alpha_j)/\sigma+(\beta_i-\beta_j)x/\sigma\}\, dv \\ &= \int_{-\infty}^{\infty} \phi(v)\, \Phi(v+A_i+B_i x) \prod_{j\neq i,\, c} \Phi\{v+(A_i-A_j)+(B_i-B_j)x\}\, dv, \end{aligned} \tag{2·1}$$

where

$$A_i = (\alpha_i-\alpha_c)/\sigma, \quad B_i = (\beta_i-\beta_c)/\sigma \qquad (i = 1, \ldots, c-1). \tag{2·2}$$

This model differs radically from other generalizations of probit analysis to more than two categories, in that the categories of response are not necessarily ordered. The model of the Aitchison & Silvey (1957) generalization visualizes c ordered categories, a series of instars or stages in the development of an insect. Independent, nonnegative, random variables $y_1, \ldots, y_{c-1}$ represent the times spent by an insect in successive instars. The response category of an insect subjected to stimulus strength x, time from hatching, is then determined as

$$\begin{aligned} &1 \quad \text{if} \quad y_1 > x, \\ &i \quad \text{if} \quad y_1+\ldots+y_{i-1} \leqslant x, \quad y_1+\ldots+y_i > x \quad (i = 2, \ldots, c-1), \\ &c \quad \text{if} \quad y_1+\ldots+y_{c-1} \leqslant x. \end{aligned}$$

The models of Ashford (1959) and Gurland, Lee & Dahm (1960) essentially use the concept of a natural tolerance y, distributed as $N(\alpha+\beta x, \sigma^2)$. This natural tolerance falls into one of the c ordered intervals $(-\infty, \gamma_1)$, (γ_1, γ_2), ..., (γ_{c-1}, ∞), thus determining the category of response. All of these previous generalizations lead to category probabilities $p_i(x)$ which can be expressed in terms of differences of standardized normal distribution functions. None of them is relevant to the applications envisaged for the present model and, therefore, we have to face up to whatever difficulties arise from the integral expression (2·1).

3. Identifiability

In this section we shall concern ourselves with the question of identifiability; a general result will be proved specifying which functions of the parameters are actually identifiable. It transpires that we are always able to estimate the polychotomous counterparts of the LD 50 strength.

First the role played by the stimulus strength in the identifiability problem must be made

apparent. The experiment of subjecting an individual to a stimulus of strength x is a binomial trial with success probability

$$p_1(x) = \Phi(A+Bx).$$

In a single binomial trial the success, or response, probability completely identifies the model, but for any specified value θ of this success probability there will correspond many (A, B), for example any A and B satisfying $A+Bx = \Phi^{-1}(\theta)$.

Hence, for the model of a single binomial trial, A and B are not identifiable. For a model, however, consisting of two independent binomial trials at different stimulus strengths x_1 and x_2, the parameters A and B are identifiable since the equations

$$\theta_1 = p_1(x_1) = \Phi(A+Bx_1),$$

$$\theta_2 = p_1(x_2) = \Phi(A+Bx_2),$$

have clearly a unique solution for A and B in terms of θ_1 and θ_2. A formal proof, by the implicit function theorem, provides a guide to the form of the proof in the general case. The Jacobian

$$\begin{aligned} J_2 &= \frac{\partial\{p_1(x_1), p_1(x_2)\}}{\partial(A,B)} \\ &= \begin{vmatrix} \phi(A+Bx_1) & x_1\phi(A+Bx_1) \\ \phi(A+Bx_2) & x_2\phi(A+Bx_2) \end{vmatrix} \\ &= (x_2-x_1)\,\phi(A+Bx_1)\,\phi(A+Bx_2) \\ &\neq 0 \quad (x_1 \neq x_2), \end{aligned} \tag{3·1}$$

and, by the implicit function theorem, the uniqueness of the solution, and hence the identifiability of A and B, is established. Thus, in our formulation, $(\alpha_1-\alpha_2)/\sigma$ and $(\beta_1-\beta_2)/\sigma$ are identifiable and form a maximum identifiable set.

The general result is contained in the following theorem.

THEOREM. *For the model consisting of two multinomial trials, the first with category probabilities $p_1(x_1), \ldots, p_c(x_1)$ and the second with category probabilities $p_1(x_2), \ldots, p_c(x_2)$, where $x_1 \neq x_2$, and $p_i(x)$ is given by* (2·1), *the parameters $A_1, \ldots, A_{c-1}, B_1, \ldots, B_{c-1}$ are identifiable.*

Proof. Identifiability will be established if we can show that

$$J_c = \frac{\partial\{p_1(x_1), \ldots, p_{c-1}(x_1), p_1(x_2), \ldots, p_{c-1}(x_2)\}}{\partial(A_1, \ldots, A_{c-1}, B_1, \ldots, B_{c-1})} \neq 0$$

for distinct x_1, x_2. Two simple relations are required:

(i) $$\partial p_i/\partial B_j = x\,\partial p_i/\partial A_j \quad (i = 1, \ldots, c-1;\ j = 1, 2); \tag{3·2}$$

(ii) $$\partial p_i/\partial A_i = d_i - \sum_{k \neq i} \partial p_i/\partial A_k \quad (i = 1, \ldots, c-1), \tag{3·3}$$

where $$d_i = \int_{-\infty}^{\infty} \phi(v)\,\phi(v+A_i+B_ix) \prod_{j \neq i} \Phi\{v+(A_i-A_j)+(B_i-B_j)x\}\,dv.$$

Thus, from (3·2),

$$J_c = \begin{vmatrix} M_1 & x_1M_1 \\ M_2 & x_2M_2 \end{vmatrix}, \tag{3·4}$$

where

$$M_i = \begin{bmatrix} \partial p_1(x_i)/\partial A_1 & \dots & \partial p_1(x_i)/\partial A_{c-1} \\ \vdots & & \vdots \\ \partial p_{c-1}(x_i)/\partial A_1 & \dots & \partial p_{c-1}(x_i)/\partial A_{c-1} \end{bmatrix} \quad (i = 1, 2). \tag{3·5}$$

Hence,

$$J_c = \begin{vmatrix} M_1 & 0 \\ M_2 & (x_2 - x_1) M_2 \end{vmatrix} = (x_2 - x_1)^{c-1} |M_1| \, |M_2|. \tag{3·6}$$

The determinants $|M_1|$ and $|M_2|$ both take the form

$$\begin{vmatrix} d_1 - \sum_{k \neq 1} m_{1k} & m_{12} & \dots & m_{1,c-1} \\ m_{21} & d_2 - \sum_{k \neq 2} m_{2k} & \dots & m_{2,c-1} \\ \vdots & \vdots & & \vdots \\ m_{c-1,1} & m_{c-1,2} & \dots & d_{c-1} - \sum_{k \neq c-1} m_{c-1,k} \end{vmatrix}, \tag{3·7}$$

where, for $i \neq k$,

$$\begin{aligned} m_{ik} &= \partial p_i(x)/\partial A_k \\ &= -\int_{-\infty}^{\infty} \phi(v) \, \phi\{v + (A_i - A_k) + (B_i - B_k) x\} \, \Phi(v + A_i + B_i x) \\ &\qquad \times \prod_{j \neq i, k, c} \Phi\{v + (A_i - A_j) + (B_i - B_j) x\} dv. \end{aligned} \tag{3·8}$$

Note that we have used (3·3) in setting

$$m_{ii} = d_i - \sum_{k \neq i} m_{ik} \quad (i = 1, \dots, c-1). \tag{3·9}$$

The nonvanishing property of the Jacobian J_c will follow if we can establish that any determinant of the above form (3·7) is nonzero. The result contained in the Appendix asserts the positivity of such determinants, and so the identifiability property is proved.

While the basic parameters $\alpha_1, \dots, \alpha_c, \beta_1, \dots, \beta_c, \sigma$ of the general model of § 2 are unidentifiable, the Theorem and the fact that functions of identifiable parameters are identifiable ensure the identifiability of all standardized differences of α's and of β's, such as

$$(\alpha_i - \alpha_j)/\sigma.$$

Critical values of the stimulus strength occur where two indicant means, for example $\alpha_i + \beta_i x$ and $\alpha_j + \beta_j x$, are equal. A typical critical strength is thus

$$\xi_{ij} = -(\alpha_i - \alpha_j)/(\beta_i - \beta_j). \tag{3·10}$$

From the above identifiability considerations we see that all ξ_{ij} are identifiable. The interpretation of ξ_{ij} in terms of category probabilities is that

$$p_i(\xi_{ij}) = p_j(\xi_{ij}). \tag{3·11}$$

Thus the LD 50 concept of probit analysis is replaced by a set of $\frac{1}{2}c(c-1)$ critical strengths. The relevance of these in the practical problem of utility estimation mentioned in § 1 is that they are crucial in determining the optimum behaviour of a reasoning subject.

4. Estimation

While identifiability of parameters is a necessary condition for their estimability it is not a sufficient condition in that there may be no absolute maximum of the likelihood function at finite values of the parameters. For example, in a simple probit analysis, observation

of the reactions of two individuals, the first submitted to strength x_1 and the second to strength $x_2 \neq x_1$, does not allow estimation of A and B. While considerable insight into the general concept of estimability might be gained from a thorough investigation of this particular case we feel that it is less important from a practical point of view. Data which lead to inestimability are generally easily discovered through a breakdown of the computational procedure of estimation. This section therefore deals only with the techniques involved in obtaining estimates of the parameters A_i and B_i $(i = 1, \ldots, c-1)$.

The procedure for $c = 2$ is well known, as this is exactly a probit situation. For the method of estimation see, for example, Finney (1947). An extension of the probit method is required before we can handle values of c in excess of 2. In fact, we have set out the procedure completely only for $c = 3$. For higher values of c, the procedure is similar, but requires either an algorithm for the evaluation of multivariate normal integrals (Plackett, 1954; Steck, 1958) to replace the bivariate algorithm used here or, more promisingly, application of recent techniques using Hermite–Gauss quadrature (Sowden & Ashford, 1969) directly to the generalized form of (4·1).

The case $c = 3$. One suitable technique is the usual maximum likelihood adaptation of Newton's method of solution for a system of equations. Studies of its application to a number of artificial situations of the type of § 5 suggest that ill-conditioning arises only when the data are sparse in relation to the inherent variability σ. This, the price of trying to do too much with too little data, must inevitably attend other methods of estimating. In the application of the technique to the study of large scale decision-making envisaged by Aitchison (1970), the repetitive nature of the decision problems will provide an adequate supply of data.

An obvious obstacle to the maximum likelihood approach is the evaluation of integrals like

$$I = \int_{-\infty}^{\infty} \phi(v)\, \Phi(v+a)\, \Phi(v+b)\, dv. \tag{4·1}$$

A polynomial approximation for Φ, suitable for use on a computer, is available, giving high speed and accuracy to seven places of decimals; see Ibbetson (1963).

Since it is easy to express (4·1) as the integral over a rectangle of a standardized bivariate normal density function, we have, in this case, preferred this approach to a quadrature technique. The required bivariate integral can be approximated very quickly on a computer to, at worst, three places of decimals (Cadwell, 1951; Owen, 1956). Writing $\phi(\cdot|\mu, \sigma^2)$ for the density function of a $N(\mu, \sigma^2)$ distribution, we have

$$\begin{aligned} \frac{\partial I}{\partial a} &= \int_{-\infty}^{\infty} \phi(v)\, \phi(v+a)\, \Phi(v+b)\, dv \\ &= \phi(a|0, 2) \int_{-\infty}^{\infty} \phi(v|-\tfrac{1}{2}a, \tfrac{1}{2})\, \Phi(v+b)\, dv \\ &= \phi(a|0, 2)\, \Phi\{(2b-a)/\sqrt{6}\}, \end{aligned} \tag{4·2}$$

by the convolution theorem. Hence

$$I = \int_{-\infty}^{a} \phi(x|0, 2)\, \Phi\{(2b-x)/\sqrt{6}\}\, dx + G(b),$$

where $G(b)$ is a function of b alone; letting $a = -\infty$, we see that $G(b) \equiv 0$. Thus

$$I = \int_{-\infty}^{a} \phi(x|0, 2)\left\{\int_{-\infty}^{(2b-x)/\sqrt{6}} \phi(y)\, dy\right\} dx,$$

which, after the substitution

$$X = x/\sqrt{2}, \quad Y = y\sqrt{3}/2 + x/(2\sqrt{2}),$$

becomes

$$\frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{a/\sqrt{2}}\int_{-\infty}^{b/\sqrt{2}} \exp\left\{\frac{-1}{2(1-\rho^2)}(X^2 - 2\rho XY + Y^2)\right\} dX\, dY = B(a/\sqrt{2}, \quad b/\sqrt{2}; \tfrac{1}{2}), \quad (4.3)$$

for $\rho = \frac{1}{2}$, in the notation of Owen (1956). Thus

$$p_1(x) = B[(A_1 + B_1 x)/\sqrt{2}, \quad \{A_1 - A_2 + (B_1 - B_2)\, x\}/\sqrt{2}; \tfrac{1}{2}], \quad (4{\cdot}4)$$

$$p_2(x) = B[(A_2 + B_2 x)/\sqrt{2}, \quad \{A_2 - A_1 + (B_2 - B_1)\, x\}/\sqrt{2}; \tfrac{1}{2}]. \quad (4{\cdot}5)$$

If we now write n_i for the total number of individuals subjected to strength x_i, and r_i and s_i for the numbers among the n_i in response categories 1 and 2, respectively, then the likelihood is effectively

$$\prod_{i=1}^{g} \{p_1(x_i)\}^{r_i}\{p_2(x_i)\}^{s_i}\{1 - p_1(x_i) - p_2(x_i)\}^{n_i - r_i - s_i},$$

g being the number of distinct x's. The log likelihood function, with subscripts removed for simplicity, is

$$L = \Sigma\{r \log p_1 + s \log p_2 + (n - r - s) \log (1 - p_1 - p_2)\}. \quad (4{\cdot}6)$$

If we write

$$\left.\begin{aligned} E_1 &= \phi(A_1 + B_1 x|0, 2)\, \Phi[\{A_1 - 2A_2 + (B_1 - 2B_2)\, x\}/\sqrt{6}], \\ E_2 &= \phi(A_2 + B_2 x|0, 2)\, \Phi[\{A_2 - 2A_1 + (B_2 - 2B_1)\, x\}/\sqrt{6}], \\ E_3 &= \phi\{A_1 - A_2 + (B_1 - B_2)\, x|0, 2\}\, \Phi[\{A_1 + A_2 + (B_1 + B_2)\, x\}/\sqrt{6}], \end{aligned}\right\} \quad (4{\cdot}7)$$

$$\left.\begin{aligned} C_1 &= (E_1 + E_3)/p_1 + E_1/(1 - p_1 - p_2), \quad C_2 = -E_3/p_2 + E_1/(1 - p_1 - p_2), \\ C_3 &= -E_3/p_1 + E_2/(1 - p_1 - p_2), \quad C_4 = (E_2 + E_3)/p_2 + E_2/(1 - p_1 - p_2), \end{aligned}\right\} \quad (4{\cdot}8)$$

then the likelihood equations, after simplification with the use of (4·2), can be expressed as

$$\left.\begin{aligned} 0 = \partial L/\partial A_1 &= \Sigma n[C_1\{(r/n) - p_1\} + C_2\{(s/n) - p_2\}], \\ 0 = \partial L/\partial B_1 &= \Sigma nx[C_1\{(r/n) - p_1\} + C_2\{(s/n) - p_2\}], \\ 0 = \partial L/\partial A_2 &= \Sigma n[C_3\{(r/n) - p_1\} + C_4\{(s/n) - p_2\}], \\ 0 = \partial L/\partial B_2 &= \Sigma nx[C_3\{(r/n) - p_1\} + C_4\{(s/n) - p_2\}]. \end{aligned}\right\} \quad (4{\cdot}9)$$

If, further, we write

$$w_1 = C_1(E_1 + E_3) - C_2 E_3, \quad w_2 = -C_1 E_3 + C_2(E_2 + E_3), \quad w_3 = -C_3 E_3 + C_4(E_2 + E_3), \quad (4{\cdot}10)$$

then the information matrix is

$$I_0 = \begin{bmatrix} \Sigma n w_1 & \Sigma nx w_1 & \Sigma n w_2 & \Sigma nx w_2 \\ \Sigma nx w_1 & \Sigma nx^2 w_1 & \Sigma nx w_2 & \Sigma nx^2 w_2 \\ \Sigma n w_2 & \Sigma nx w_2 & \Sigma n w_3 & \Sigma nx w_3 \\ \Sigma nx w_2 & \Sigma nx^2 w_2 & \Sigma nx w_3 & \Sigma nx^2 w_3 \end{bmatrix}. \quad (4{\cdot}11)$$

The iterative method of solution is given by

$$I_{\theta}\begin{bmatrix} A_{1j}-A_{1,j-1} \\ B_{1j}-B_{1,j-1} \\ A_{2j}-A_{2,j-1} \\ B_{2j}-B_{2,j-1} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial A_1} \\ \frac{\partial L}{\partial B_1} \\ \frac{\partial L}{\partial A_2} \\ \frac{\partial L}{\partial B_2} \end{bmatrix}, \tag{4·12}$$

where A_{1i} is the ith iterate of A_1, and similarly for B_1, A_2 and B_2. The initial approximations A_{10}, B_{10}, A_{20} and B_{20} can conveniently be set at zero.

The use of the inverse of (4·11) as the variance matrix involves the usual invocation of asymptotic properties of maximum likelihood estimators, and the supposition that we have sufficient data to justify this asymptotic assumption. A main aim is not so much a study of the accuracy of the estimates obtained as the ability of the fitted model to simulate the envisaged subject. We have, therefore, not felt compelled at this stage to make simulation studies to discover the adequacy of asymptotic theory for particular n.

5. TEST DATA

The data for testing purposes was constructed as follows, for the case $c = 3$.

We chose the values $\alpha_1 = 10{\cdot}4, \beta_1 = 0{\cdot}01, \alpha_2 = 9{\cdot}0, \beta_2 = 0{\cdot}04, \alpha_3 = 4{\cdot}0, \beta_3 = 0{\cdot}10$ and $\sigma = 1{\cdot}0$. One hundred values of the stimulus strength x were drawn from the uniform distribution on the selected interval (0, 120), using tables of random numbers. For each x, tables of random normal deviates permitted the construction of a randomized y_i from the $N(\alpha_i + \beta_i x, \sigma^2)$ distribution ($i = 1, 2, 3$): the greatest of these y_i ($i = 1, 2, 3$) determined the category corresponding to the value of x used. Figure1 shows the lines $y = \alpha_i + \beta_i x$ ($i = 1, 2, 3$) in the range considered.

With the initial approximations $A_{10} = B_{10} = A_{20} = B_{20} = 0$, the KDF9 computer at Glasgow University, operating under the Egdon system, took less than forty-five seconds to produce answers, which are shown in Table 1 together with the chosen values of the parameters.

Table 1. *Calculated and chosen values of the parameters*

Parameters	Calculated values	Chosen values
A_1	7·907	6·400
B_1	−0·108	−0·090
A_2	6·423	5·000
B_2	−0·078	−0·060
ξ_{12}	48·889	46·667
ξ_{13}	73·297	71·111
ξ_{23}	82·855	83·333

The variance-covariance matrix of the estimators is

$$\begin{bmatrix} 2{\cdot}598 & -0{\cdot}032 & 2{\cdot}431 & -0{\cdot}028 \\ -0{\cdot}032 & 0{\cdot}000 & -0{\cdot}029 & 0{\cdot}000 \\ 2{\cdot}431 & -0{\cdot}029 & 2{\cdot}435 & -0{\cdot}028 \\ -0{\cdot}028 & 0{\cdot}000 & -0{\cdot}028 & 0{\cdot}000 \end{bmatrix}.$$

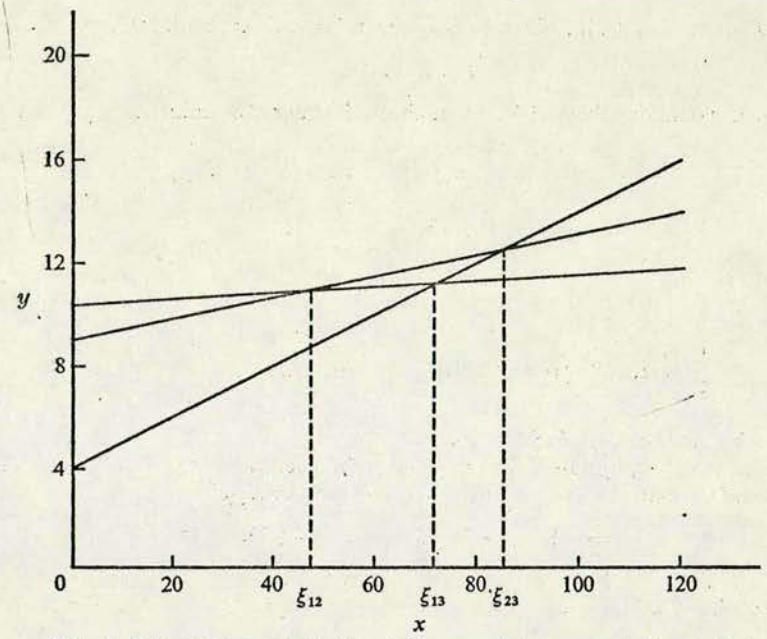

Fig. 1. The three mean indicant lines used to generate the test data.

Appendix

We now prove the positivity of the determinant Δ_n ($n \geqslant 2$), defined by

$$\Delta_n = \begin{vmatrix} d_1 - \sum_{k \neq 1} m_{1k} & m_{12} & \dots & m_{1n} \\ m_{21} & d_2 - \sum_{k \neq 2} m_{2k} & \dots & m_{2n} \\ \vdots & \vdots & & \vdots \\ m_{n1} & m_{n2} & \dots & d_n - \sum_{k \neq n} m_{nk} \end{vmatrix},$$

where $d_i > 0$ ($i = 1, \dots, n$) and $m_{ij} < 0$ ($i \neq j$).

The method used is induction on n. Specifically, we take as our induction assumption: $\Delta_n > 0$ ($n = 2, \dots, N$). We then prove that $\Delta_{N+1} > 0$.

By adding all other columns to the first, we see that

$$\Delta_{N-1} = \begin{vmatrix} d_1 & m_{12} & \dots & m_{1,N+1} \\ d_2 & d_2 - \sum_{k \neq 2} m_{2k} & \dots & m_{2,N+1} \\ \vdots & \vdots & & \vdots \\ d_{N+1} & m_{N+1,2} & \dots & d_{N+1} - \sum_{k \neq N+1} m_{N+1,k} \end{vmatrix},$$

that is,

$$\Delta_{N+1} = d_1 D_{11} + (-1) m_{12} D_{12} + (-1)^2 m_{13} D_{13} + \dots + (-1)^N m_{1,N+1} D_{1,N+1}, \tag{A 1}$$

where D_{ij} is the minor of the (i, j)th element in Δ_{N+1}. Now

$$D_{11} = \begin{vmatrix} d_2' - \sum_{\substack{k=2 \\ k \neq 2}}^{N+1} m_{2k} & m_{23} & \dots & m_{2,N+1} \\ m_{32} & d_3' - \sum_{\substack{k=2 \\ k \neq 3}}^{N+1} m_{3k} & \dots & m_{3,N+1} \\ \vdots & \vdots & & \vdots \\ m_{N+1,2} & m_{N+1,3} & \dots & d_{N+1}' - \sum_{\substack{k=2 \\ k \neq N+1}}^{N+1} m_{N+1,k} \end{vmatrix},$$

where $d_i' = d_i - m_{i1}$ $(i = 2, \ldots N+1)$. This is an Nth order determinant and, by the induction assumption,

$$D_{11} > 0. \tag{A2}$$

By subtracting all other columns from the first in the Nth order determinant D_{12} we have

$$D_{12} = \begin{vmatrix} d_2 - \sum_{\substack{k=2 \\ k \neq 2}}^{N+1} m_{2k} & m_{23} & \cdots & m_{2,N+1} \\ m_{31}+m_{32} & d_3 - \sum_{\substack{k=1 \\ k \neq 3}}^{N+1} m_{3k} & \cdots & m_{3,N+1} \\ \vdots & \vdots & & \vdots \\ m_{N+1,1}+m_{N+1,2} & m_{N+1,3} & \cdots & d_{N+1} - \sum_{\substack{k=1 \\ k \neq N+1}}^{N+1} m_{N+1,k} \end{vmatrix},$$

which, by the induction assumption, is positive.

It is not difficult to see that, by $(i-2)$ interchanges of adjacent rows in D_{1i} $(i = 3, \ldots, N+1)$, D_{1i} may be expressed in a form analogous to the above expression for D_{12}. Hence

$$(-1)^i D_{1i} > 0 \quad (i = 2, \ldots, N+1). \tag{A3}$$

Thus, finally, by (A1), (A2) and (A3) it follows that $\Delta_{N+1} > 0$.

This completes the induction step, and it remains only to prove the result for $n = 2$. We have

$$\begin{aligned} \Delta_2 &= \begin{vmatrix} d_1 - m_{12} & m_{12} \\ m_{21} & d_2 - m_{21} \end{vmatrix}, \\ &= d_1 d_2 - d_1 m_{21} - d_2 m_{12} \\ &> 0. \end{aligned}$$

REFERENCES

AITCHISON, J. (1970). Statistical problems of treatment allocation. *J. R. Statist. Soc.* A. **133** (to appear).

AITCHISON, J. & SILVEY, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika* **44**, 131–43.

ASHFORD, J. R. (1959). An approach to the analysis of data for semiquantal responses in biological assay. *Biometrics* **15**, 573–81.

CADWELL, J. H. (1951). The bivariate normal integral. *Biometrika* **38**, 475–9.

FINNEY, D. J. (1947). *Probit Analysis*. Cambridge University Press.

GURLAND, J., LEE, I. & DAHM, P. A. (1960). Polychotomous quantal response in biological assay. *Biometrics* **16**, 382–98.

IBBETSON, D. (1963). Gauss. *Communications Ass. Computing Machinery* **6**, 616.

OWEN, D. B. (1956). Tables for computing bivariate normal probabilities. *Ann. of Math. Statist.* **27**, 1075–90.

PLACKETT, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika* **41**, 351–60.

SOWDEN, R. R. & ASHFORD, J. R. (1969). Computation of the bi-variate normal integral. *Appl. Statist.* **18**, 169–80.

STECK, G. P. (1958). A table for computing trivariate normal probabilities. *Ann. Math. Statist.* **29**, 780–800.

[*Received November* 1969. *Revised February* 1970]

9 STATISTICAL PREDICTION ANALYSIS

Following the recognition of the central role of predictive density functions in all prediction problems, and particularly at the stage of development in 1965 it was natural to seek formulations of some standard problems in terms of the new approach. In the three publications associated with his Ph.D. thesis therefore Dunsmore (1966, 1968, 1969) explored the reformulation of the problems of classification, calibration, regulation and optimisation as decision problems, albeit with simple and perhaps oversimplified utility functions, involving predictive aspects. These, together with a realisation that in an increasing number of these consultative problems predictive methods could give conclusions radically different from other popular and strongly advocated methods, led to the conviction that some major work explaining, developing and popularising predictive methods should be undertaken. This conviction resulted in the publication of Aitchison and Dunsmore (13:1975), and it is to the developments reported there that we now turn. Further strong motivation for the book is to be found expressed in its preface.

The aim of the book was to develop the ideas and the methodology in such a way that they were seen as an extremely tractable tool by the user. The developments over previous work already quoted are as follows.

1. An extension of the treatment of predictive density functions, with, in particular, a full account of the multinormal situation, leading to d-dimensional Student, Siegel and Student-Siegel distributions for future mean, covariance matrix, and

joint (mean, covariance matrix) distributions.

2. A full and usable explanation of predictive regression analysis.

3. A full study of decisive precision problems relating design values associated with all-or-nothing, linear-loss and quadratic-loss utility structures to the mode, percentiles and mean of the predictive distribution.

4. A more thorough study of the predictive density function as a means of conveying information about prediction. In particular, most plausible Bayesian prediction intervals, regions of previous experience and atypicality indices are new concepts developed.

5. A new expository device is introduced for the description of tolerance regions, namely the coverage distributions, the distributions of $P_f\{R(x)|\theta\}$ induced by the distributions $P_e(x|\theta)$. This concept allows the simple definition of c-mean coverage and (c,g) guaranteed coverage tolerance regions in terms of the mean and percentiles of these distributions. Mean-coverage and guaranteed coverage intervals for the two-parameter negative-exponential distribution are provided for the first time.

6. Sampling inspection is presented as a predictive decision problem, with a utility function of the form U(a,y), where y is the relevant characteristic of a typical item of output and a a possible action, such as scrap the untested items. For particular specifications of U(a,y) mean coverage and guaranteed coverage predictors are seen to be appropriate, and so circumstances under which there is support for such predictors are identified.

7. The nature of a calibration or assay problem is spelt out carefully in terms of model assumptions concerning standards and specimens and the relation of specimens to standards. Unlike the decision theory specification of Dunsmore (1968) this approach leads to the notion of the calibrative distribution, expressing for each specimen with given response the relative probabilities of the possible calibrative values. The method is compared favourably with a number of other approaches to the problem.

8. A similar spelling out of the modelling assumptions is used in a study of the relevance of predictive distributions to the problem of medical diagnosis. In this approach diagnosis is regarded as an essentially inference problem, with aim the supply of realistic probabilities for the possible disease types rather than a decision problem involving allocation to a particular type. The reason for this is the almost invariable absence of anything approaching a realistic utility function. The favourite object of study, the overall misclassification rate, is included in this condemnation as being possibly misleading, and in some actual cases ridiculous as a measure of quality of diagnosis. Thus the approach adopted provides the clinician at the end of the diagnostic phase of patient management with as realistic odds as possible leaving entirely to him, in the absence of a properly quantified utility structure, the informal process of deciding on treatment in the light of the diagnostic assessments provided. The assumptions lead to the use of predictive distributions in deriving the crucial tool, the diagnostic

distribution. For the multivariate normal cases this leads to the predictive method suggested by Geisser (1964).

What no one seemed to have realised up to this point in development is how radically different this predictive statistical diagnosis can be from the popular (for example, BMD and other computer packages) estimative statistical diagnosis methods. By estimative diagnosis we mean the simple plugging in of estimates in the model as if they were the true values, in sharp contrast to the weighting process according to reliability inherent in the predictive method. Aitchison and Dunsmore (1975) provide a real example, from a Glasgow hospital, where the odds assigned by the two methods are so at variance as to suggest completely different treatments. This difference will be discussed in more detail in the next section.

A further feature introduced here is the provision of atypicality indices, on a scale 0 (absolutely typical) to 1 (completely atypical), with respect to each of the disease types as a system of monitoring whether a new case may have wandered into the wrong clinical network. If all the atypicality indices are close to 1 then such a suspicion is aroused. Again radical differences between estimative and predictive assessments of atypicality indices emerge.

AITCHISON, J. and DUNSMORE, I.R. (1975)

Statistical Prediction Analysis

Cambridge Univeristy Press

A copy of this book is reproduced in volume 2

10 ESTIMATIVE AND PREDICTIVE MODEL FITTING

'When two statistical methods applied to the same important practical problem provide answers of such enormous difference that they can commonly lead to radically different practical consequences - even to the difference between curing and harming a patient - it is high time to subject them to a critical comparison. Such is the current situation with two methods of statistical discrimination distinguished by the terms *estimative* and *predictive*.' So ran the introductory remarks, based on experience of real medical diagnostic applications, of Aitchison, Habbema and Kay (17:1977). We now examine the relative merits of these methods as claimants to realism.

Statisticians often refer to the process of 'fitting the model' $p(y|\theta)$ for the experiment f, say, from the information x arising from an experiment e, described by density function $p(x|\theta)$ on X. The intention of the process often seems to be the assessment of the whole density function $p(y|\theta)$, and certainly in some applications such as calibration and diagnosis this is an appropriate aim. A popular way of fitting the model is first to estimate θ by $\hat{\theta}(x)$, for example a maximum likelihood estimate, and then to regard

$$p\{x|\theta = \hat{\theta}(x)\}$$

as the fitted density function. This process of replacing parameter by estimate has been termed the estimative method of model-fitting (Aitchison and Dunsmore, 13:1975). It is often supposed to be a satisfactory procedure (Boneva, Kendall and Stefanov, 1971) when there appears to be no specific purpose to the exercise such as

hypothesis testing. Aitchison and Dunsmore (13:1975, p.228) point out that the procedure really must be suspect since it is a case of putting all one's eggs in one basket, namely $\hat{\theta}(x)$, without taking any account of the unreliability of $\hat{\theta}$ as an estimator.

The predictive method of model-fitting uses as an assessment of $p(y|\theta)$ a predictive density function:

$$p(y|x) = \int_{\Theta} p(y|\theta)p(\theta|x)d\theta.$$

In the construction of $p(y|x)$ we are not picking out one particular value of θ, such as $\hat{\theta}(x)$ in the estimative method, but weighting the possible $p(y|\theta)$ by $p(\theta|x)$, this weight being an assessment of the probability of θ on the basis of x and the prior $p(\theta)$. Even if $p(\theta)$ is not too well specified we might prefer $p(y|x)$ to $p\{y|\theta = \hat{\theta}(x)\}$, since any reasonable weighting should be more sensible than ignoring unreliability of estimates altogether. Thus on purely intuitive grounds we should expect the predictive method to make better sense than the estimative method.

When $p(y|\theta)$ is of $N(\mu,\sigma^2)$ form and e is n replicates of f, with the usual estimates $\bar{x}$ and s of μ and σ, we have as estimative, fitted model a $N(\bar{x},s)$ form and as predictive fitted model, on the basis of the standard vague prior, $St\{n-1,\bar{x},(1+n^{-1})s^2\}$ in the notation of Aitchison and Dunsmore (13:1975). If we were dealing with a problem of estimation or hypothesis-testing we would instruct our students to use the St distribution rather than the N distribution. Why then should the situation be different if the purpose is different from straightforward estimation or hypothesis-testing?

Aitchison (14:1975) provides further theoretical support for the use of predictive fitting in preference to estimative fitting.

Since, at least in some applications, the objective is to obtain a good assessment of the true $p(y|\theta)$ by a fitted density function, say $q(y|x)$, we should perhaps judge success by some measure of closeness of the shot $q(y|x)$ to the target $p(y|\theta)$. One such measure, well based in information theory, is the Kullback-Liebler (1951) directed divergence

$$D(p,q) = \int_Y p(y|\theta) \log \frac{p(y|\theta)}{q(y|x)} dy.$$

If $r(y|x)$ is a rival to $q(y|x)$ then $q(y|x)$ is 'closer' to p than $r(y|x)$ if

$$M(p:q,r) = D(p;r) - D(p;q)$$

$$= \int_Y p(y|\theta) \log \frac{q(y|x)}{r(y|x)} dy$$

is positive. This measure depends on x and so we are forced to assess the relative merits of q and r as methods of fitting $p(y|\theta)$ by considering their relative performance in repeated applications against a background of replication of e, that is in terms of the criterion

$$\int_X p(x|\theta)dx \int_Y p(y|\theta) \log \frac{q(y|x)}{r(y|x)} dy .$$

Aitchison (14:1975) shows that if $p(y|\theta)$ is multivariate normal $N_d(\mu,\Sigma)$, if $r(y|x)$ is taken to be the estimative fit $N_d(\bar{x},s)$ and $q(y|x)$ to be the predictive form $St_d\{n-1,\bar{x},(1+n^{-1})S\}$ then the above criterion is positive. Thus on this criterion the estimative fit is inferior to the predictive fit based on the vaguest prior distribution (Aitchison and Dunsmore, 13:1975, Table 2.3). See also Murray (1977) for an optimum property of this choice of prior.

We reemphasise that this inferiority of the estimative fit to the predictive fit is being assessed on a criterion which makes no assumption of knowledge of a prior distribution.

A similar advantage of predictive over estimative fit is established for gamma modelling.

Apart from these theoretical considerations there is another effective way, namely simulation, of examining the greater claim to realism in multivariate normal modelling of the predictive method. This is a main objective of Aitchison, Habbema and Kay (17:1977) within the context of statistical diagnosis.

It is relatively easy to simulate d-dimensional normal vectors from any distribution. Suppose then that we simulate n_1 vectors from a known $N_d(\mu_1,\Sigma_1)$ distribution, n_2 vectors from a known $N_d(\mu_2,\Sigma_2)$ distribution. We can then, using these as a diagnostic training set, construct both an estimative and a predictive diagnostic system, based say on a prevalence ratio of π_1/π_2. Suppose that the odds assigned by the estimative and predictive diagnostic methods (Aitchison and Dunsmore, 13:1975, Chapter 11) for a new case with feature vector y are $q_1(y)/q_2(y)$ and $r_1(y)/r_2(y)$. Since we know the simulative distributions we have an absolute standard, namely

$$\frac{p_1(y)}{p_2(y)} = \frac{\pi_1\ \phi_d(y|\mu_1,\Sigma_1)}{\pi_2\ \phi_d(y|\mu_2,\Sigma_2)},$$

against which to judge the quality of the estimative and the predictive values.

Aitchison, Habbema and Kay (17:1977) carry out a number of such simulations and find overwhelming support in favour of the predictive method in these comparisons. Since the predictive

distributions are just as easy to compute as the estimative distributions there seems little excuse for not applying the predictive method.

It is interesting to comment here that the predictive method, which has a Bayesian origin although it can equally be regarded as a simple weighting device, seems to have sparked off a response by frequentist supporters. For example, Moran and Murphy (1979) adapt the estimative approach to diagnosis so that it more nearly conforms to the predictive approach. It is difficult to resist making a cynical comment here on fashions in statistical theory. There was a time when it was fashionable for Bayesians to devise Bayesian methods which conformed with long-used frequentist methods (Lindley, 1965). Now it seems that some frequentists, appreciating the effectiveness of some Bayesian methods, are determined to show that the Bayesian results can equally be established through frequentist arguments.

Having thus established the relevance of predictive density functions in straightforward parametric statistical modelling we can now turn to specific applications. First, in §11 we use predictive diagnosis as a norm against which to judge subjective diagnostic judgments; and secondly, in §12 we consider the use of predictive parametric modelling in more complex diagnostic situations.

AITCHISON, J. (1975)

Goodness of prediction fit

Reprinted from *Biometrika* 62, 547-54

Biometrika (1975), 62, 3, *p.* 547
With 3 text-figures
Printed in Great Britain

Goodness of prediction fit

BY J. AITCHISON
Department of Statistics, University of Glasgow

SUMMARY

Fitting a parametric model or estimating a parametric density function plays an important role in a number of statistical applications. Two widely-used methods, one replacing the unknown parameter by an efficient estimate and so termed estimative and the other using a mixture of the possible density functions and commonly termed predictive, are compared. On a general criterion of closeness of fit based on a discriminating information measure the predictive method is shown to be preferable. Explicit measures of the relative closeness of predictive and estimative fits are obtained for gamma and multinormal models.

Some key words: Estimation of density functions; Goodness of fit; Predictive distributions.

1. INTRODUCTION

Suppose that a class of parametric models with sample space Y, parameter set Θ and class

$$\mathscr{P} = \{p(y \mid \theta) : \theta \in \Theta\}$$

of density functions on Y, is postulated for some observational situation f. Suppose further that we have available data x from some experiment e, for example n replicates of f, described by a class of similarly parameterized models $\{X, \Theta, p(x \mid \theta)\}$. The 'fitting' of a model for f or the 'estimation' of the true density function $p(y \mid \theta)$ on the basis of the data x involves the choice of a density function, $q(y \mid x)$ say, from some class $\mathscr{Q} \supseteq \mathscr{P}$ of density functions on Y. The intention is clearly that $q(y \mid x)$ should in some sense be close to the true density function $p(y \mid \theta)$, but in a variety of statistical activities where this idea is explicitly or implicitly used the criterion of closeness is seldom defined. The purpose of this paper is to examine the consequences of using a familiar measure of overall closeness, the Kullback & Liebler (1951) discriminating information measure, and in particular to compare from this viewpoint two different methods of fitting models.

The class $\mathscr{Q}$ may be identical with $\mathscr{P}$, for example, when we fit the model by setting

$$q(y \mid x) = p\{y \mid \theta = \hat{\theta}(x)\} \quad (y \in Y), \tag{1·1}$$

where $\hat{\theta}(x)$ is some efficient estimate of θ. Suppose, however, that we set

$$q(y \mid x) = \int_{\Theta} p(y \mid \theta)\, p(\theta \mid x)\, d\theta \quad (y \in Y), \tag{1·2}$$

where $p(\theta \mid x)$ can be regarded simply as a weighting function based on the data x or a Bayesian posterior density function for θ based on a prior $p(\theta)$ and the data x. Here $\mathscr{Q}$ is larger than $\mathscr{P}$, a feature which may simply recognize that closeness to the true density function is a more compelling consideration in some applications than an insistence that the estimating density function belongs to $\mathscr{P}$.

Although (1·1) and (1·2) are not the only ways of fitting a model they are two quite different methods which are current competitors in a number of areas of application. To distinguish clearly between the two methods we term (1·1) an estimative density function, since the true density function is estimated by the insertion of an estimate for the parameter; and we use the term predictive density function for (1·2) because of its already established role and terminology in statistical prediction theory.

The rivalry between these two methods occurs in goodness-of-fit testing (Guttman, 1967; Hager & Antle, 1968) and in a variety of direct prediction problems where probabilistic statements about a future observation from f are to be made on the basis of the data x from e (Aitchison & Dunsmore, 1975, p. 1). Much of this rivalry stems from differing assessments of appropriateness of classical and Bayesian approaches to specific applications. There are, however, important situations where some measure of overall closeness of the fitted density function to the true density function is a help in deciding which of the two methods is more realistic in its approach. For example, in the practical operation of a statistical discrimination or diagnostic system based on data on individuals of known types and with observed feature vectors, the assignment of type probabilities to a new individual of unknown type on the basis of his observed feature vector y is of primary importance. Irrespective of which method is adopted this requires estimates of likelihoods and so of the true probability density associated with each type. Applied to a number of new individuals with different feature vectors this technique effectively calls for the estimation of the true feature vector density functions for each type.

Within this specific application Anderson (1958, p. 137) and the widely applied computer package described by Dixon (1970) use the estimative method, whereas Geisser (1964) and Aitchison & Kay (1975) use the predictive method. That the practical consequences of these different methods can be enormous has been shown by Aitchison & Kay (1975). It is therefore of considerable interest to pose the question as to which method yields the more realistic type probabilities. One purpose of the present paper is to explain why these results could be expected on theoretical grounds by a comparison of the relative merits of estimative and predictive density functions as estimates of the true density function. Note that we are here concerned with the provision of realistic type probabilities, for example in clinical medicine as a diagnostic guide to the appropriate allocation of treatment to a patient, and not in the assessment of realistic misclassification probabilities as considered, for example, by Lachenbruch & Mickey (1968).

As far as asymptotic properties of the methods are concerned it is obvious that under very general conditions (1·1) tends to the true density function, and Geisser (1971) has established a number of similar consistency properties for the corresponding predictive form (1·2). For large samples, therefore, it is clear that the difference between estimative and predictive density functions will be of little practical importance. We therefore emphasize that the criterion we investigate in this paper is concerned with properties of samples of finite size.

2. A GOODNESS-OF-FIT CRITERION

In attempting to judge the goodness of fit of $q(y \mid x)$ to the unknown $p(y \mid \theta)$ we require some overall measure of the divergence of $q(y \mid x)$ from the true $p(y \mid \theta)$. Since the true $p(y \mid \theta)$ is our target an appropriate measure is the Kullback & Leibler (1951) directed measure of divergence

$$D\{p(y \mid \theta), q(y \mid x)\} = \int_Y p(y \mid \theta) \log \left\{\frac{p(y \mid \theta)}{q(y \mid x)}\right\} dy, \tag{2·1}$$

which is positive unless $q(y \mid x)$ coincides with $p(y \mid \theta)$.

If we have two contenders, say $q(y \mid x)$ and $r(y \mid x)$, for the role of estimate of $p(y \mid \theta)$ then $q(y \mid x)$ is closer than $r(y \mid x)$ if

$$\begin{aligned} M(p; q, r) &= D(p, r) - D(p, q) \\ &= \int_Y p(y \mid \theta) \log \left\{\frac{q(y \mid x)}{r(y \mid x)}\right\} dy \end{aligned} \tag{2·2}$$

is positive. This measure depends on θ and the particular x observed. If we are to avoid Bayesian arguments, at least for the present, then we are forced to attempt to assess the relative merits of q and r as methods of estimating p by considering their relative performance in repeated applications against a background of replication of e. The long-run measure of relative closeness will then be represented by the expectation of M with respect to $p(x \mid \theta)$:

$$\int_X p(x \mid \theta)\, dx \int_Y p(y \mid \theta) \log \left\{\frac{q(y \mid x)}{r(y \mid x)}\right\} dy. \tag{2·3}$$

This still suffers from the drawback that in general it will depend on θ but we shall see that there are important cases where (2·3) is independent of θ, in which case it provides a powerful criterion of closeness.

Let us proceed with the general case and follow through formally the consequences of imbedding the estimation of $p(y \mid \theta)$ in a sequence of recurring problems where nature produces θ according to a density function $p(\theta)$ and where the informative experiment e yields x with density function $p(x \mid \theta)$. The natural measure of relative closeness is then the expectation of (2·3) with respect to $p(\theta)$ giving

$$\int_\Theta p(\theta)\, d\theta \int_X p(x \mid \theta)\, dx \int_Y p(y \mid \theta) \log \left\{\frac{q(y \mid x)}{r(y \mid x)}\right\} dy. \tag{2·4}$$

Since $p(\theta)\, p(x \mid \theta) = p(x)\, p(\theta \mid x)$ we can express (2·4), after a change of order of integration, as

$$\int_X p(x)\, dx \int_Y p(y \mid x) \log \left\{\frac{q(y \mid x)}{r(y \mid x)}\right\} dy, \tag{2·5}$$

where

$$p(y \mid x) = \int_\Theta p(y \mid \theta)\, p(\theta \mid x)\, d\theta, \tag{2·6}$$

the predictive density function described in §1 and based on prior $p(\theta)$ and data x. It follows immediately that on the basis of criterion (2·4) the predictive density function (2·6) is unrivalled in its closeness to $p(y \mid \theta)$. For taking $q(y \mid x) = p(y \mid x)$ gives as inner integral

in (2·5) a Kullback & Liebler directed measure $D\{p(y \mid x), r(y \mid x)\}$, and this is positive for any $r(y \mid x)$ different from $p(y \mid x)$. Hence (2·5), and so (2·4), is positive.

The predictive density function $p(y \mid x)$ is the natural Bayesian method of estimating the true density function. All we have so far established in the above optimality property of the predictive density function is the tautology that when faced with an essentially Bayesian situation, with given $p(\theta)$, it is good sense to act in a Bayesian way. We shall see in the next section, however, that we can exploit this result to obtain an interesting extension of the good sense of predictive density function estimation.

3. Inadequacy of estimative fit

Although the result of the previous section demonstrates the superiority of the predictive density function over all other contenders as a fit to the class of models when a specific prior distribution is known it gives no guarantee that some other density function may not be superior when no prior distribution can be assumed. In the absence of a prior distribution it is tempting to use the simply constructed estimative density function $p\{y \mid \theta = \hat{\theta}(x)\}$ as the fitted model. We can show, however, that in two important practical situations such a procedure is inadequate as measured by goodness-of-fit criterion (2·3). More specifically, given an estimative density function, say $r(y \mid x)$, based on data x, we can construct another density function $q(y \mid x)$, interpretable as a predictive density function, superior to $r(y \mid x)$ in terms of the positivity for all θ of the goodness-of-fit criterion (2·3). We emphasize that this criterion makes no assumption about a prior distribution so that the density function $q(y \mid x)$, although formally constructed by way of a $p(\theta)$, is competing against $r(y \mid x)$ on the latter's terms.

Gamma case. We suppose that the class of models to be fitted is $\mathrm{Ga}(K, \theta)$, that is with

$$p(y \mid \theta) = \theta^K y^{K-1} e^{-\theta y}/\Gamma(K) \quad (y > 0),$$

with K known and θ the indexing parameter. Suppose further that the informative experiment, possibly summarized by an appeal to sufficiency arguments, can be described by a density function $p(x \mid \theta)$ which is $\mathrm{Ga}(k, \theta)$.

The maximum likelihood estimate based on data x is $\hat{\theta}(x) = k/x$, so that as estimative density function we take

$$r(y \mid x) = \frac{k^K y^{K-1}}{x^K \Gamma(K)} \exp\left(-\frac{ky}{x}\right). \tag{3·1}$$

For a prior density function $p(\theta)$ of $\mathrm{Ga}(g, h)$ form, the posterior density function $p(\theta \mid x)$ is $\mathrm{Ga}(G, H)$ with

$$G = g + k, \quad H = h + x, \tag{3·2}$$

and the corresponding predictive density function (2·6) is

$$p(y \mid x) = \frac{H^G}{B(K, G)} \frac{y^{K-1}}{(H+y)^{K+G}} \quad (y > 0), \tag{3·3}$$

an inverse beta distribution, written $\mathrm{Inbe}(K, G, H)$, say. Suppose that we compare $r(y \mid x)$ with any predictive density function of the form (3·3) with $h = 0$ and $g \geqslant 0$ so that $G \geqslant k$: in other words, we take

$$q(y \mid x) = \frac{x^G}{B(K, G)} \frac{y^{K-1}}{(x+y)^{K+G}} \quad (y > 0). \tag{3·4}$$

Then

$$\log\left\{\frac{q(y\,|\,x)}{r(y\,|\,x)}\right\} = \log\left\{\frac{\Gamma(K+G)}{\Gamma(G)}\right\} - K\log k + k\frac{y}{x} - (K+G)\log\left(1+\frac{y}{x}\right). \tag{3·5}$$

We note that this expression depends on x and y only through the ratio y/x. The distribution of $z = y/x$, for given θ, is Inbe $(K, G, 1)$ and so is independent of θ, say with density function $p(z)$. Since $E(z) = K/(G-1)$ and since

$$\int_0^\infty p(z)\log(1+z)\,dz = \psi(K+G) - \psi(G), \tag{3·6}$$

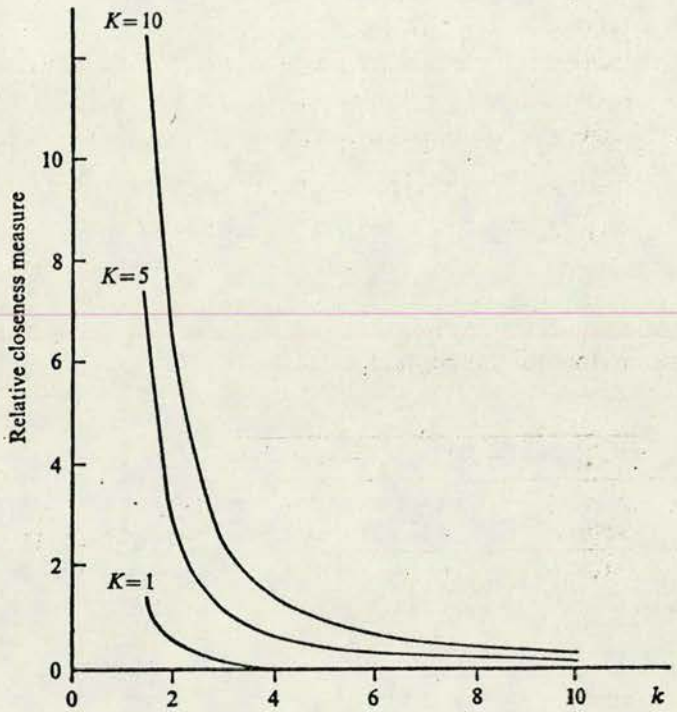

Fig. 1. Relative closeness measures for the gamma case, with $K = 1, 5, 10$.

where $\psi(G) = \Gamma'(G)/\Gamma(G)$ is the digamma function tabulated by Abramowitz & Stegun (1964, p. 267), we can easily express criterion (2·3) in the following form:

$$W(K, G, k) = \log\left\{\frac{\Gamma(K+G)}{\Gamma(G)}\right\} - K\log k + \frac{Kk}{G-1} - (K+G)\{\psi(K+G) - \psi(G)\}. \tag{3·7}$$

Since (3·7) does not depend on θ its value is unaltered by multiplying by the $p(\theta)$ associated with $q(y\,|\,x)$ and integrating over Θ. This leads us from (2·3) to a criterion of the form (2·5) which we know to be positive. Thus (3·7) is positive and we note the inadequacy of the estimative fit relative to such a predictive form. The argument is, of course, hardly rigorous since the multiplying $p(\theta)$ is an improper prior, but the positivity of (3·7) could be established directly. We show in Fig. 1 the graphs of (3·7) against k for the cases $K = 1$, i.e. the exponential distribution, 5, 10 and for the 'vaguest' of priors, for which $g = 0$ and

so $G = k$. The superiority of the predictive over the estimative form is appreciable for k/K small. If k is large relative to K then (3·7) is small. This is obvious since, for fixed K,

$$q(y \mid x) \sim r(y \mid x) \quad (k \to \infty),$$

and the increase in k can be interpreted in practical situations as an increase in the 'size' of the informative experiment e.

Multinormal case. The argument for the multinormal case runs very close to that for the gamma case. The class of models to be fitted is d-dimensional multinormal $N_d(\mu, \Sigma)$ with

$$p(y \mid \mu, \Sigma) = (2\pi)^{-\frac{1}{2}d} |\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(y-\mu)' \Sigma^{-1}(y-\mu)\}. \tag{3·8}$$

We can suppose that the description of the informative experiment can be conveniently condensed into the observation of $x = (m, S)$, where m and S are effective mean and covariance matrix estimators distributed with normal and Wishart forms with effective sample size n and effective degrees of freedom ν. Then the estimative fit is

$$r(y \mid x) = N_d(m, S) \tag{3·9}$$

and predictive fits based on 'vaguest' prior information (Aitchison & Dunsmore, 1975, p. 21) take the form

$$q(y \mid x) = \mathrm{St}_d[\nu, m, \{1 + (1/n)\} S], \tag{3·10}$$

where the d-dimensional vector z is $\mathrm{St}_d(k, b, c)$ if it has density function

$$\frac{\Gamma\{\frac{1}{2}(k+1)\}}{\pi^{\frac{1}{2}d}\Gamma\{\frac{1}{2}(k-d+1)\} |kc|^{\frac{1}{2}} \{1 + (z-b)'(kc)^{-1}(z-b)\}^{\frac{1}{2}(k+1)}}. \tag{3·11}$$

Here

$$\log\left\{\frac{q(y \mid x)}{r(y \mid x)}\right\} = \log\left[\frac{\Gamma\{\frac{1}{2}(\nu+1)\}}{\Gamma\{\frac{1}{2}(\nu-d+1)\}}\right] - \tfrac{1}{2}d \log\left\{\tfrac{1}{2}\nu\left(1+\frac{1}{n}\right)\right\} + \tfrac{1}{2}\nu\left(1+\frac{1}{n}\right) z - \tfrac{1}{2}(\nu+1)\log(1+z), \tag{3·12}$$

where

$$z = (y-m)'(\nu S)^{-1}(y-m)/\{1 + (1/n)\}. \tag{3·13}$$

The fact that (3·13) depends on m and S only through z and that z has an

$$\mathrm{Inbe}\{\tfrac{1}{2}d, \tfrac{1}{2}(\nu-d+1), 1\}$$

distribution, irrespective of μ and Σ, allows the evaluation

$$E\left[\log\left\{\frac{q(y \mid x)}{r(y \mid x)}\right\}\right] = W\left\{\tfrac{1}{2}d, \tfrac{1}{2}(\nu-d+1), \tfrac{1}{2}\nu\left(1+\frac{1}{n}\right)\right\} \tag{3·14}$$

in terms of the function defined in (3·7). The positivity of (3·14) is then established by an argument similar to that for the gamma case.

We show in Fig. 2 the graphs of (3·14) against ν for dimensions $d = 1, 4, 8$ and with $n = \nu + 1$, corresponding to the case where m and S are estimates of μ and Σ, both based on n replicates of the multinormal experiment. Clearly as the dimensionality of the multinormal distribution increases the more suspect the use of the estimative method becomes unless the extent ν of the experimentation is sufficiently large to make the difference

between $q(y \mid x)$ and $r(y \mid x)$ small. In the more general case where ν and n are not directly related, for fixed d and ν the value of (3·14) increases as n decreases.

We reemphasize that the inferiority of the estimative fit to a predictive fit is being assessed on a criterion, namely (2·3), which makes no assumption of knowledge of a prior distribution. For $d \geqslant 3$ it could, of course, be argued that the use of m as an estimate of μ may already be suspect on admissibility grounds (Stein, 1956) and that for $d \geqslant 3$ we ought to be comparing some adjusted form of (3·9) against a correspondingly adjusted predictive

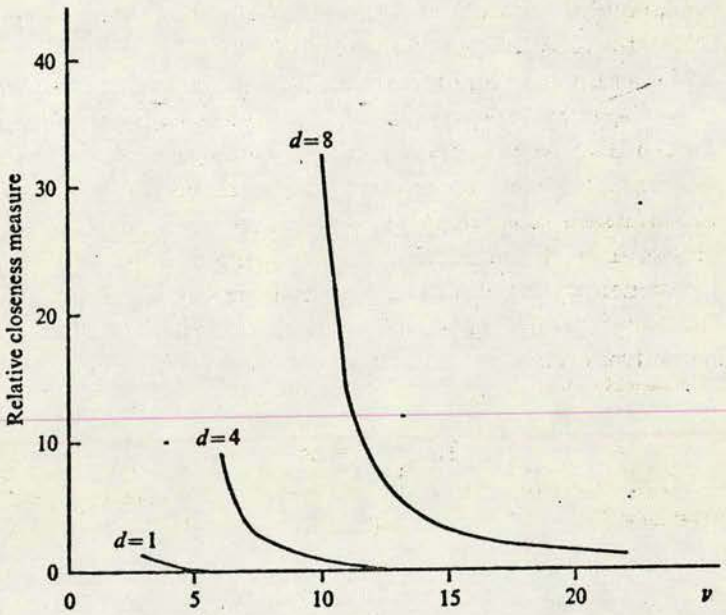

Fig. 2. Relative closeness measures for the multinormal case, with $d = 1, 4, 8$ and $\nu = n-1$.

form. We have not attempted to do this although we might conjecture results similar to those already found simply because the predictive form will again retain the advantage of taking account of sampling variability. The admissibility argument, however, is concerned with the direct estimation of μ, and appropriate loss structures for this are of doubtful relevance to the problem of density function fit.

4. Discussion

In their paper on nonparametric problems of estimating density functions Boneva, Kendall & Stefanov (1971, p. 3), in a passing remark on parametric estimation, say that for estimation of a univariate normal density function the obvious method is to estimate the two parameters in the standard manner. Although this may be so if the estimated density function is restricted to the given parametric class there is no need for such a restriction in many practical problems. From the viewpoint of obtaining a close overall fit to the true density function we have seen that such replacement of the parameters by estimates can be inadequate compared with adopting some suitable mixture of the basic models as is achieved by the predictive function. It can, of course, be argued that the use of the estimative density function (1·1) is perfectly sound provided that we take account of the

sampling variability of $\hat{\theta}(x)$ by providing a confidence interval, say, for each density $p(y \mid \theta)$ or probability $\mathrm{pr}(A \mid \theta)$. Unfortunately the proviso is all too easily ignored. When θ appears in a complicated way in the density function, as for example as mean vector and covariance matrix in a multinormal density function, the temptation to resort to the estimative density function, as by Dixon (1970), can be overwhelming, whereas the multivariate Student form of the predictive density function is easily computed.

We have demonstrated the Kullback & Liebler superiority of the predictive method for only two members of the exponential family, the gamma and normal. For the other main members, the binomial and the Poisson, we cannot even define the measure of divergence (2·1) for the estimative fits since the measure is infinite.

The Kullback & Liebler measure is only one possible measure of relative closeness of two competing estimates of density functions. For other measures the superiority of the predictive density function may well be reduced or altogether overthrown. For example, if θ is a mean parameter and the measure places emphasis on good estimation of densities in the neighbourhood of the mean the estimative could easily prove superior to the predictive method. But the lesson from the analysis of this paper must be to avoid the too facile insertion of $\hat{\theta}(x)$ in the density function as an obvious way of obtaining an estimate. The precise purpose of the density function estimation has to be considered and some appropriate measure of closeness introduced.

REFERENCES

ABRAMOWITZ, M. & STEGUN, I. A. (1964). *Handbook of Mathematical Functions*. New York: Dover.

AITCHISON, J. & DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

AITCHISON, J. & KAY, J. W. (1975). Principles, practice and performance in decision making in clinical medicine. In *Proc. 1973 NATO Conference on the Role and Effectiveness of Decision Theory*. London: English Universities Press.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

BONEVA, L. I., KENDALL, D. G. & STEFANOV, I. (1971). Spline transformations: three diagnostic aids for the statistical data-analyst. *J. R. Statist. Soc.* B **33**, 1–71.

DIXON, W. J. (1970). *BMD Biomedical Computer Programs*. University of California Press.

GEISSER, S. (1964). Posterior odds for multivariate normal classifications. *J. R. Statist. Soc.* B **26**, 69–76.

GEISSER, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference*. Ed. V. P. Godambe and D. A. Sprott, pp. 459–69. Toronto: Holt, Rinehart and Winston.

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. R. Statist. Soc.* B **29**, 82–100.

HAGER, H. & ANTLE, C. (1968). The choice of the degree of a polynomial model. *J. R. Statist. Soc.* B **30**, 469–71.

KULLBACK, S. & LIEBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 525–40.

LACHENBRUCH, P. A. & MICKEY, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1–10.

STEIN, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp.* **1**, 197–206.

[*Received March* 1975. *Revised June* 1975]

AITCHISON, J., HABBEMA, J.D.F. and KAY, J.W. (1977)

A critical comparison of two methods of statistical discrimination

Reprinted from *Applied Statistics* 26, 15-25

Appl. Statist. (1977),
26, *No.* 1, p. 15.

A Critical Comparison of Two Methods of Statistical Discrimination

By J. AITCHISON J. D. F. HABBEMA and J. W. KAY

University of Glasgow, Britain *Universities of Rotterdam and Leiden, Holland* *University of Glasgow, Britain*

[Received May 1976. Revised September 1976]

SUMMARY

Important clinical differences arising in the application of commonly advocated discriminant or diagnostic methods demand a thorough assessment of the realism of their different assessments. Recent theoretical work on the estimation of density functions provides reasons for these differences and suggests which methods should provide greater realism. These suggestions are strongly supported by a simulation study. Specific recommendations are made concerning statistical diagnostic practice.

Keywords: ATYPICALITY INDEX; DISCRIMINANT ANALYSIS: ESTIMATION OF DENSITY FUNCTION; ESTIMATIVE DIAGNOSIS; INFORMATION DIVERGENCE MEASURE; PREDICTIVE DIAGNOSIS; SIMULATION

1. PRACTICAL DIFFERENCES BETWEEN TWO METHODS

WHEN two statistical methods applied to the same important practical problem provide answers of such enormous difference that they can commonly lead to radically different practical consequences—even to the difference between curing and harming a patient—it is high time to subject them to a critical comparison. Such is the current situation with two methods of statistical discrimination which we shall presently distinguish by the terms *estimative* and *predictive*. The purpose of this paper is to explain the reasons for these differences, to pose the question of which method is likely to yield the more reliable results and to go some way in answering that important question.

These practical differences are well illustrated by the problem of differential diagnosis of Conn's syndrome and have been reported by Aitchison and Dunsmore (1975, p. 231), Aitchison and Kay (1975). Their comparisons are summarized and extended in Figs 1 and 2. Conn's syndrome is a rare condition producing high blood pressure and is now known to have two quite different causes: (1) a benign tumour in one adrenal gland, curable by surgical removal, or (2) a more diffuse condition affecting both adrenal glands with the possibility of control of blood pressure by drug treatment. Accurate diagnosis of type can only be achieved by microscopic examination of adrenal tissue removed at an operation. Since for most patients with type 2, surgery is inadvisable clinicians faced with a new patient known to have Conn's syndrome obviously require a realistic preoperative assessment of the relative plausibilities of the two types in order to help in their difficult treatment decision (Brown *et al.*, 1971). For this purpose only a small *basic* set of past records consisting of 20 confirmed cases of type 1 and 11 cases of type 2 is available (Aitchison and Dunsmore, 1975, Table 1.6) with eight measured aspects for each case: age, plasma concentrations of sodium, potassium, bicarbonate, renin and aldosterone; systolic and diastolic blood pressures.

Because of positive skewness in the raw data logarithmic transformations were carried out and subsequently tests of multivariate normality for each type along the lines of Andrews *et al.* (1973) were conducted. No significant departures from multinormality were detected. Although—as in all goodness-of-fit testing of parametric models and particularly for small data sets—lack of significance cannot be construed as actual support for the model, we follow

common practice in adopting a multinormal assumption for the transformed data of each type in what follows. Tests of comparison of some of the marginal variances for the two types do, however, show significant differences and these suggest that this multinormal assumption should allow for possible inequality of the two covariance matrices.

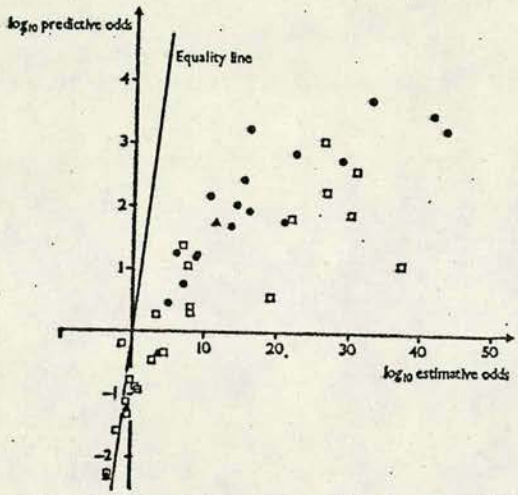

FIG. 1. Comparison of log-odds as assessed by estimative and predictive methods. ●, New case of type 1; ▲, new case of type 2; □, new case of unknown type.

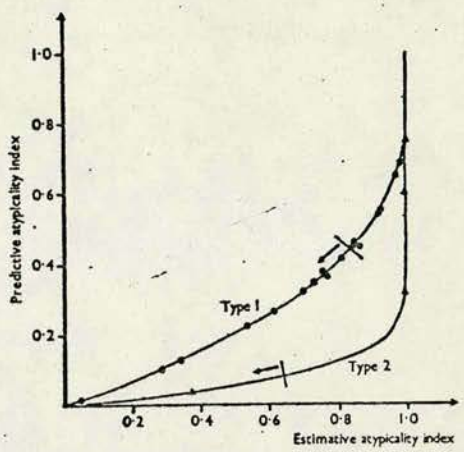

FIG. 2. Comparison of atypicality indices as assessed by estimative and predictive methods. Arrows indicate indices "within previous experience". ●, New case of type 1; ▲, new case of type 2.

We suppose that prior to observation of a new patient there is an agreed assessment of the relative plausibilities of the two possible types. This would normally be taken to be the incidence rate at the clinic under consideration and would be based on previous experience in the clinic. The differences that arise between the two methods are in no way associated with different assessments of incidence and it is therefore convenient to compare the methods from a common starting point, say equal incidence rates for the two types.

Under these identical assumptions of equal incidence and multinormality, after transformation, of observation vectors two methods of obtaining the required relative plausibility assessments are available. We may follow the *estimative* method of, for example, Anderson (1958, p. 137) and Dixon (1970) in replacing parameters in appropriate likelihood ratios by their estimates. Or we may use the *predictive* method of Geisser (1964), Dunsmore (1966) and Aitchison and Kay (1975) which replaces likelihood ratios by ratios of predictive density functions. These two methods are more fully described in Section 2; our immediate interest is to compare their practical effect.

Fig. 1 shows, for each of a test set of 43 new cases, the plot of predictive assessment of $\log_{10}$ (odds on type 1) against the corresponding estimative assessment, together with the accurate diagnosis for the 21 cases where treatment by surgery has allowed confirmation. The fewness of the confirmed cases of type 2 is due to surgery being contraindicated for most of such cases. The vast differences between the predictive and estimative assessments of odds are obvious and can affect decision about treatment.

Both methods will, of course, assign plausibilities for a new case, even if it has been wrongly referred or if its observation vector contains some gross inaccuracy. A sensible precaution in the application of any diagnostic system is therefore the monitoring of new cases to ensure that they are not too atypical of all the possible types considered. For this purpose each case can be assigned atypicality indices for each of the possible types (Aitchison and Kay, 1975). The observation vector of a given case will lie on a particular ellipsoid of concentration of the

underlying multinormal distribution of observation vectors associated with the type. An atypicality index of the case for the type is then simply an assessment of the probability content of the interior of this ellipsoid. Thus atypicality indices are measured on the interval (0, 1); the nearer the index is to 1 the more atypical is the case. If a case has all its atypicality indices near 1 then wrong referral or faulty measurement of features must be seriously considered.

For a particular type atypicality indices assessed by the two methods are mathematically related and so their joint variation can be represented by a locus in the unit square. Fig. 2 shows this locus for each of the types, 1 and 2. We can also show on this diagram for each type the most atypical cases in the basic set of 31 cases, the position of the point with greatest atypicality index denoting the limit of "previous experience". Note that there is a remarkable difference between the estimative and predictive assessments. For example, for type 2 the proportion of new cases we may expect within previous experience is on a predictive assessment 0·08, whereas the estimative method puts this expectancy as high as 0·64. The corresponding proportions for type 1 are 0·44 and 0·84. For the 21 new cases for which the accurate diagnosis is known the atypicality indices relative to the known type are shown. Of the 4 new type 2 cases three are assigned type 2 atypicality indices of almost 1 by the estimative method. Of the complete test set of 43 new cases, known to have Conn's syndrome in some form, the estimative method assesses that a high proportion, 14 out of 43, have both atypicality indices greater than 0·95, whereas the predictive method has no such doubly atypical case in its assessments.

The differences in odds and atypicality indices assigned by the two methods clearly indicate that it is important to come to some verdict as to which are the more realistic.

2. Description of the Two Methods

In its simplest form statistical diagnosis or discrimination is concerned with assessing to which of a given finite set T of possible *types* an individual or *case* belongs on the basis of a vector $\mathbf{x}$ of observations on that individual. Often for the user, as in the clinical situation of Section 1, such assessments are most conveniently expressed by the statistician in terms of the probabilities or plausibilities of the possible types, rather than the choice of a specific type; and we shall consider this assessment of plausibilities the aim of *statistical diagnosis*.

For statistical diagnosis to be useful the distribution of observation vectors must differ from type to type. Suppose that some parametric form $p(x|t, \theta)$ can be assumed for the probability (density) function of x for given type t, with the parameter θ usually multidimensional. For example, for the multinormal situation of Section 1, $\theta = (\mu_1, \mu_2, \Sigma_1, \Sigma_2)$, where μ_1 and μ_2 are the vector means and Σ_1 and Σ_2 the covariance matrices associated with the two types. If we knew the true value of the parameter, say θ, there would be no problem and no controversy. For a given incidence rate $p(t)$ the probability $p(t|x, \theta)$, that we would assign to type t for a case with observation vector $\mathbf{x}$, would be computed from Bayes' formula as

$$p(t|\mathbf{x}, \theta) \propto p(t)p(\mathbf{x}|t, \theta), \tag{2.1}$$

where the $\propto$ sign indicates that the scaling factor required to obtain equality does not depend on the argument t of the probability function. In practice we never know θ but we usually have available data z, say, from a basic set of past case records consisting of observation vectors on cases of known types. The differences between the methods arise from different ways in which we make inferences about θ from z and how we use these inferences in effect to replace $p(\mathbf{x}|t, \theta)$ on the right side of (2.1) to obtain a plausibility assessment $p(t|\mathbf{x}, z)$ for the type of a new case with observation vector $\mathbf{x}$.

A popular method is to proceed as if these distributions were known, with the parameter value θ replaced by some efficient estimate $\hat{\theta}(z)$, often a maximum likelihood estimate. Thus $p(\mathbf{x}|t, \theta)$ on the right side of (2.1) is replaced by

$$r(\mathbf{x}|t, z) = p\{\mathbf{x}|t, \theta = \hat{\theta}(z)\}. \tag{2.2}$$

For such a method which places emphasis on first obtaining estimates of the unknown parameters we use the term *estimative* diagnosis. Such methods are to be found in the Biomedical Data Processing package (Dixon, 1970), in the commonly applied likelihood ratio techniques (Anderson, 1958, p. 133); for example, the linear discriminant technique is often justified on these grounds (Anderson, 1958, p. 137).

A more recent and radically different method replaces $p(\mathbf{x}|t,\theta)$ on the right side of (2.1) by

$$q(\mathbf{x}|t,z) = \int_{\Theta} p(\mathbf{x}|t,\theta)p(\theta|z)\,d\theta, \tag{2.3}$$

where $p(\theta|z)$ can be regarded either as some weighting function based on the data z or as a full Bayesian posterior density function for θ based on a prior $p(\theta)$ and on the data z. The form (2.3) commonly arises in Bayesian statistical prediction theory as the predictive density function for a "future" observation $\mathbf{x}$ on a case of type t as assessed on the basis of the data z; see, for example, Jeffreys (1961), Geisser (1964), Guttman and Tiao (1964), Aitchison and Sculthorpe (1965) and Zellner and Chetty (1965). Because of this association the method has come to be known as *predictive* diagnosis (Geisser, 1964; Dunsmore, 1966; Aitchison and Dunsmore, 1975; Aitchison and Kay, 1975).

The nature of the difference between the two methods can be well illustrated by the case where $p(\mathbf{x}|t,\theta)$ is multinormal, as in the medical diagnostic application of this paper. Suppose that $p(\mathbf{x}|t,\theta)$ is a d-dimensional multinormal density function with mean μ_t and covariance matrix Σ_t: we then write

$$p(\mathbf{x}|t,\theta) = N_d(\mu_t,\Sigma_t) = (2\pi)^{-\frac{1}{2}d}|\Sigma_t|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\mathbf{x}-\mu_t)'\Sigma_t^{-1}(\mathbf{x}-\mu_t)\}. \tag{2.4}$$

We collect here for convenience of reference only the essential results required later; for details and generalizations see Aitchison and Kay (1975). We suppose that the past case records contain n_t of type t with observation vectors $\mathbf{x}_1,\ldots,\mathbf{x}_{n_t}$, and that the prior $p(\theta)$ from which the predictive density function is constructed is the vague prior used by Aitchison and Kay (1975).

If we write m_t and S_t for the mean and covariance matrix for the sample $\mathbf{x}_1,\ldots,\mathbf{x}_{n_t}$, then

$$r(\mathbf{x}|t,z) = N_d(m_t,S_t), \tag{2.5}$$

whereas

$$q(\mathbf{x}|t,z) = St_d\left\{\nu_t, m_t, \left(1+\frac{1}{n_t}\right)S_t\right\}, \tag{2.6}$$

a d-dimensional Student-type density function, where $\nu_t = n_t - 1$, and $St_d(\nu,b,c)$ is defined on $X = R^d$ by the density at $\mathbf{x}$:

$$\frac{\Gamma\{\frac{1}{2}(\nu+1)\}}{\pi^{\frac{1}{2}d}\Gamma\{\frac{1}{2}(\nu-d+1)\}|\nu c|^{\frac{1}{2}}}\frac{1}{\{1+(\mathbf{x}-b)'(\nu c)^{-1}(\mathbf{x}-b)\}^{\frac{1}{2}(\nu+1)}}. \tag{2.7}$$

Thus the picture of (2.2) and (2.3) is of two distributions both centred on the same vector mean m_t and with the same class of ellipsoids of concentration but with the $q(\mathbf{x}|t,z)$ less concentrated than $r(\mathbf{x}|t,z)$ about m_t.

The relationship of (2.5) to (2.6) is of course the multivariate counterpart of the use of a univariate normal approximation instead of the exact t distribution in such familiar areas as significance tests and confidence intervals for normal means. Whereas in the univariate case the distinction between the two methods is often of little practical consequence even for moderately sized samples the distinction in the multivariate case can be of considerable practical significance, as has already been noted in Section 1. The fact that the two methods (2.5) and (2.6) are asymptotically equivalent for large basic sets should not delude the user of discriminant analysis into the supposition that in his practical "finite" situation there is no substantial difference between them.

If in the analysis it is assumed that the covariance matrices Σ_1 and Σ_2 are equal then (2.5) and (2.6) become

$$r(\mathbf{x}|t,z) = N_d(m_t, S), \tag{2.8}$$

$$q(\mathbf{x}|t,z) = St_d\left\{\nu, m_t, \left(1+\frac{1}{n_t}\right)S\right\}, \tag{2.9}$$

where

$$\nu = n_1+n_2-2, \quad \nu S = \nu_1 S_1+\nu_2 S_2.$$

The use of (2.5), (2.6), (2.8) and (2.9) as estimates of $p(\mathrm{v}|t,\theta)$ are thus four different methods of statistical diagnosis which we may conveniently denote by *Pu*, *Eu*, *Pe* and *Ee* where *E* and *P* refer to the methods, estimative and predictive; and *e* and *u* relate to the assumption about covariance matrices, equal or unequal.

A convenient way to record the quantitative difference between (2.5) and (2.6) is the following:

$$\log\frac{q(\mathbf{x}|t,z)}{r(\mathbf{x}|t,z)} = \log\frac{\Gamma(\frac{1}{2}n_t)}{\Gamma\{\frac{1}{2}(n_t-d)\}} - \tfrac{1}{2}\log\frac{n_t^2-1}{2n_t} + \tfrac{1}{2}w_t(\mathbf{x}) - \tfrac{1}{2}n_t\log\left(1+\frac{n_t w_t(\mathbf{x})}{n_t^2-1}\right), \tag{2.10}$$

where

$$w_t(\mathbf{x}) = (\mathbf{x}-m_t)' S_t^{-1}(\mathbf{x}-m_t) \tag{2.11}$$

is the Mahalanobis distance with respect to the type t basic set.

We say that a case with observation vector $\mathbf{y}$ is more typical of type t than a case with observation vector $\mathbf{x}$ if

$$p(\mathbf{y}|t,\theta) > p(\mathbf{x}|t,\theta).$$

The set $R_t(\mathbf{x})$ of all observation vectors more typical of t than observation vector $\mathbf{x}$ is then

$$R_t(\mathbf{x}) = \{\mathbf{y}: p(\mathbf{y}|t,\theta) > p(\mathbf{x}|t,\theta)\}. \tag{2.12}$$

An index of atypicality for t of a case with observation vector $\mathbf{x}$ may then be defined as the assessed probability $I_t(\mathbf{x})$ of obtaining an observation vector more typical than $\mathbf{y}$. Thus

$$I_t(\mathbf{x}) = P\{R_t(\mathbf{x})|t,\theta\}. \tag{2.13}$$

The assessments of $I_t(\mathbf{x})$ associated with the estimative and predictive distributions (2.5) and (2.6) are respectively

$$\Gamma\{\tfrac{1}{2}d; \tfrac{1}{2}w_t(\mathbf{x})\} \tag{2.14}$$

and

$$B\left(\tfrac{1}{2}d, \tfrac{1}{2}(n_t-d); \frac{w_t(\mathbf{x})}{w_t(\mathbf{x})+(n_t^2-1)/n_t}\right), \tag{2.15}$$

where $w_t(\mathbf{x})$ is given by (2.11), B denotes the incomplete beta function defined by

$$B(a,b;c) = \int_0^c u^{a-1}(1-u)^{b-1}\,du/B(a,b) \quad (0\leqslant c\leqslant 1) \tag{2.16}$$

and tabulated in Pearson (1934), and Γ the incomplete gamma function defined by

$$\Gamma(a;c) = \int_0^c u^{a-1}c^{-u}\,du/\Gamma(a) \quad (c\geqslant 0) \tag{2.17}$$

and tabulated in Pearson (1922).

3. Theoretical Aspects of the Comparison

The magnitude of the differences between the estimative and predictive odds obtained in Section 1 can be readily explained in terms of expression (2.10). Fig. 3 shows the graphs of

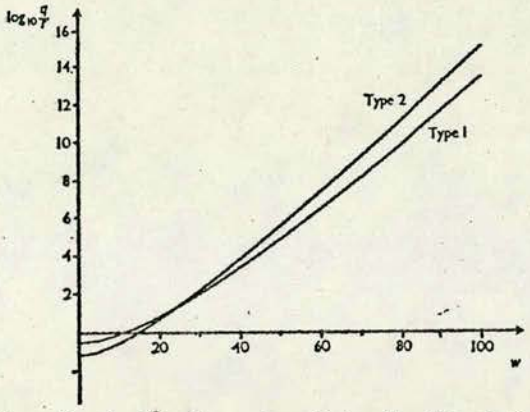

Fig. 3. Logarithm of the ratio of estimative and predictive densities plotted against Mahalanobis distance.

(2.10) plotted against w for $d = 8$, $n = 20$ and for $d = 8$, $n = 11$, the configurations of the type 1 and type 2 basic sets, respectively. The ratio $q(\mathbf{x}|t,z)/r(\mathbf{x}|t,z)$ of the predictive to the estimative assessment of the true $p(\mathbf{x}|t,\theta)$ can thus vary from approximately 10^{-1} to 10^{15} over the range $0 \leqslant w \leqslant 100$. Consider a recent new case with $w_1 = 21{\cdot}2$ and $w_2 = 73{\cdot}8$. Since the values of $\log_{10}(q/r)$ are approximately 0·94 and 9·95 at $w_1 = 21{\cdot}2$ and $w_2 = 73{\cdot}8$ respectively on the type 1 and type 2 graphs the proportional difference between the estimative and predictive odds is a factor of 10^9. Such proportional factors are common since for most cases the Mahalanobis distances w_1 and w_2 are appreciably different.

Since the evaluation of atypicality indices essentially involves integration of either $q(\mathbf{x}|t,z)$ or $r(\mathbf{x}|t,z)$ over appropriate regions it is clear from these graphs how huge discrepancies can also occur between estimative and predictive atypicality index assessments. For example, for the new case just considered the atypicality index with respect to type 2 is, for the predictive approach, by (2.15), 0·76. Since for $w_2 \geqslant 73{\cdot}8$ the estimative density is less than 10^{-9} times the predictive density the corresponding estimative assessment of atypicality index must be very close to 1.

Although the graphs of Fig. 3 provide an explanation of how the differences arise, they do nothing to resolve the question of which method is to be preferred. Estimative methods can clearly be criticized on the grounds of ignoring sampling variability of $\hat{\theta}$: if another basic set were available we would not necessarily obtain the same estimates of the parameters and some allowance ought to be made for this element of unreliability. Occasionally some lip-service is paid to this feature, for example by the provision of standard errors of scores when using the linear discriminant. On the whole, however, the problem is ignored, particularly when, as in the case of Conn's syndrome where covariance matrices are unequal, the sampling distribution theory of the crucial estimated likelihood ratio

$$q(\mathbf{x}|t=1,z)/q(\mathbf{x}|t=2,z)$$

proves intractable. The predictive method in a sense takes account of the unreliability of any estimate of θ by giving each θ an appropriate weight in forming the mixture or predictive distribution (2.3), and from a sampling variability point of view the predictive method has thus some advantage over the estimative method.

The main question of practical importance remains as yet unanswered. Which of the two methods is more realistic or gets nearer to the truth in the plausibilities it quotes? An answer based on theoretical grounds can be provided by the following argument. We recall from Section 2 that for a new case with observation vector $\mathbf{x}$ the main statistical problem is the assessment of the ratio of the true densities $p(\mathbf{x}|t,\theta)$ for $t=1,2$. The quality of this ratio is, however, strongly dependent on the quality of the individual density estimates. Since in the construction of a diagnostic system we envisage its application to a number of new cases with different observation vectors we are essentially faced with the problem of estimating whole density functions. In the assessment of atypicality indices we are directly concerned with estimating density functions.

Aitchison (1975) has considered the relative merits of the predictive and estimative methods as methods of density function estimation by defining an appropriate measure of the relative overall closeness of $q(\mathbf{x}|t,z)$ and $r(\mathbf{x}|t,z)$ to the true $p(\mathbf{x}|t,\theta)$. This measure, based on an information divergence measure of Kullback and Liebler (1951), is

$$\int_z p(z|\theta)\,dz \int_x p(\mathbf{x}|t,\theta)\log\frac{q(\mathbf{x}|t,z)}{r(\mathbf{x}|t,z)}\,d\mathbf{x},$$

the expectation of (2.10) with respect to the true density function $p(\mathbf{x}|t,\theta)$, further averaged over basic sets z. For $p(\mathbf{x}|t,\theta)$ multinormal Aitchison (1975) shows that this measure, with $q(\mathbf{x}|t,z)$ and $r(\mathbf{x}|t,z)$ assigned by (2.6) and (2.5), is independent of θ and positive, indicating greater overall closeness of the predictive density function estimator to the true density function. For the type 1 and type 2 configurations these measures with logarithms to base 10 are respectively 0·58 and 14 corresponding to the large proportional differences we have already observed in the ratio of the odds assigned by the two methods. The largeness of the measure for the type 2 group arises from the fact that 11 observation vectors of dimension 8 are an extremely small sample in multivariate terms.

Thus if the statistician wishes to have some measure of confidence in the reality of the plausibilities that he reports for the types he would be well advised on theoretical grounds to use predictive plausibilities.

4. Simulative Aspects of the Comparison

We pointed out in Section 2 that for given parametric forms for the observation vector density function $p(\mathbf{x}|t,\theta)$ the source of the statistical problem in diagnosis is our lack of knowledge of θ. If we knew θ then for a case with observation vector $\mathbf{x}$ we would know the true or realistic plausibilities we should assign to the types, namely those computed by (2.1). In any investigation by simulation we must ourselves choose θ in order to generate observation vectors; and so for any new case with observation vector $\mathbf{x}$ we have available the "true" plausibilities against which to compare those assessed by any other method based on the simulated basic set. As indicated in Section 1 we can confine attention to diagnostic situations with equal incidence rates.

For a simulation study, we consider a diagnostic situation with two types and where the vector distributions are d-dimensional multinormal $N_d(\mu_1,\Sigma_1)$ and $N_d(\mu_2,\Sigma_2)$. Since we can find a non-singular linear transformation L such that $L\Sigma_1 L' = I_d$, $L\Sigma_2 L' = \Delta$, a diagonal matrix, we can clearly, by considering the transformation $y = L(\mathbf{x}-\mu_1)$, confine simulation without loss of generality to two distributions of the forms $N_d(0,I_d)$ and $N_d(\alpha,\Delta)$. On an assumption of equality of Σ_1 and Σ_2 we have further, without loss of generality, $\Delta = I_d$. The generation of observation vectors from these distributions is then a simple routine matter. Generating n_1 and n_2 basic records of types 1 and 2, and N_1 and N_2 test records of types 1 and 2, provides a simulation situation which can be characterized by the set

$$(d,\alpha,\Delta,n_1,n_2,N_1,N_2).$$

In a simulation comparison of the four statistical diagnostic methods *Pe*, *Ee*, *Pu*, *Eu*, it is important to differentiate between simulation situations with equal covariance matrices ($\Delta = I$) and situations with unequal covariance matrices ($\Delta \neq I$). Table 1 provides details of

TABLE 1

Simulated situations included in the comparison of diagnostic methods

d, α, Δ configurations	
Equal covariance matrices, $\Delta = I_d$	Unequal covariance matrices
$\alpha = 0$, $d = 1, \ldots, 9$	$\alpha = 0$, $\Delta = 4I_d$, $d = 1, \ldots, 5$
$\alpha = e$, the vector of units, $d = 1, 2, 3$	$\alpha = e$, $\Delta = 4I_d$, $d = 1, \ldots, 9$
$\alpha_1 = 2$, $\alpha_2 = \ldots = \alpha_d = 0$, $d = 1, \ldots, 9$	$\alpha = 0$, $\Delta = 16I_d$, $d = 1, 2, 3$

Sizes of basic and test sets

d	(n_1, n_2)	d	(N_1, N_2)
1	(2, 2), (3, 3), (5, 5), (10, 10), (2, 5)	1, ..., 9	(5, 5), (10, 10)
2	(3, 3), (5, 5), (9, 9), (10, 10), (12, 12), (3, 10)		
3	(5, 5), (9, 9), (10, 10), (12, 12)		
4	(5, 5), (10, 10), (14, 14), (18, 18)		
5	(10, 10), (20, 20)		
6, ..., 9	(10, 10)		

the simulated situations which have been used to assess the relative merits of the four diagnostic methods *Pe*, *Ee*, *Pu*, *Eu*. Since substantive differences between estimative and predictive methods are known on theoretical grounds to be small for large basic sets we have concentrated attention on basic sets which are small in relation to the parameter dimension.

Each simulated training set gives rise to five assessments of log-odds in favour of type 1: the "true" one, and one for each of the four diagnostic methods. We can now compare these log-odds (*LO*) assessments, both for the basic records and also for the test records associated with the simulation situation.

The measure of performance of method i (with $i = Ee, Eu, Pe, Pu$) that we adopt for the basic set is the mean absolute deviation from the true log-odds:

$$MAD(i) = \frac{1}{n_1 + n_2} \Sigma |LO(i) - LO(\text{true})|,$$

summation being over the basic records, with a similar measure for the test set.

The total number of simulation situations considered is 110; 68 with equal covariance matrices and 42 with unequal covariance matrices. The four methods are pairwise compared for each simulation situation, both for basic records and for test records. The results are given in Table 2. The mean absolute deviation, averaged over all situations, is given for each method in Table 3.

For simulated situations with equal covariance matrices the following conclusions can be drawn:

(a1) *Pe* is by far superior to the other methods. Only in a few situations were better results obtained by *Ee* and *Pu*.

(a2) *Eu* is by far inferior to the other three methods. For no situation was the result better than for any of the other methods.

(a3) *Pu* and *Ee* are comparable both in the number of situations in which one method comes out better than the other (*Pu*: 33, *Ee*: 35) and in the magnitude of the averaged mean absolute deviations from the true log-odds.

TABLE 2

Number of simulated situations for which mean absolute deviation of row (column) method was smaller than column (row) method with e type situations in upper right triangle and u type situations in lower left triangle

Row method	*Nature of set*	*Column method*			
		Pe	*Ee*	*Pu*	*Eu*
Pe	Basic	—	65½ (2½)	62 (6)	68 (0)
	Test		67 (1)	67 (1)	68 (0)
Ee	Basic	19 (22)	—	34½ (33½)	68 (0)
	Test	12½ (29½)		35 (33)	68 (0)
Pu	Basic	24 (18)	28 (14)	—	68 (0)
	Test	29½ (12½)	34 (8)		68 (0)
Eu	Basic	2 (40)	3 (39)	2½ (39½)	—
	Test	2½ (39½)	3 (39)	1 (41)	

TABLE 3

Mean absolute deviation of log-odds for four diagnostic methods averaged over all simulated situations

Simulation type	*Nature of set*	*Meihod*			
		Pe	*Ee*	*Pu*	*Eu*
e	Basic	1·72	3·48	3·51	>100
	Test	1·80	3·70	3·36	>100
u	Basic	4·09	4·65	4·14	>100
	Test	4·43	5·44	4·27	>100

For simulated situations with unequal covariance matrices the following conclusions can be drawn:

(b1) *Pu* yields the best results, but only slightly better than *Pe* for the basic set.
(b2) *Ee* yields results that are worse than the *Pe* results.
(b3) *Eu* gives by far the worst results of all four methods.

Our overall conclusions can therefore be summarized as follows.

(i) *Eu* yields much worse results than the three other methods, even for situations involving unequal covariance matrices.

(ii) The predictive method is superior to the estimative method for both types of simulation. The *Pe* method is by far superior for situations involving equal covariance matrices. For situations involving unequal covariance matrices the *Pu* method is superior, though less markedly so.

The conclusions are of course only valid for the range of situations considered here: an effectively small number of past records from approximately multinormally distributed observations.

5. DISCUSSION

When a diagnostic or discriminant situation requires the statistician to provide plausibility assessments for the possible types we have seen that differences of practical importance can

occur between estimative and predictive methods. The theoretical considerations of Section 3 and the simulative studies of Section 4 strongly support the use of the predictive method when the feature distributions can be transformed to multinormality. We can make this recommendation more specific as follows.

When there is a high probability that the covariance matrices may differ appreciably use the *Pu* diagnostic method; otherwise use the *Pe* method.

For other distributional forms further work is clearly necessary before any general use of predictive methods could be advocated.

A common feature of the differences we have observed in applications to Conn's syndrome and other disease complexes is that predictive plausibilities are usually much closer to the equiplausible assessment than are the estimative plausibilities. It is an interesting phenomenon that while diagnosticians, such as clinicians, tend to act conservatively and underuse the information or data available to them (Taylor *et al.*, 1970) the estimative methods advocated by many statisticians tend to read too much into the data by presenting obviously extravagant odds. The source of the estimative overstatements is undoubtedly the reluctance to take account of the sampling variability problem. It is interesting to observe that it is through an essentially Bayesian approach that we have been able to overcome the sampling variability problem, usually regarded as a distinctly frequentist problem.

The improper prior, which we have used in the example we have considered, is not an essential feature of the predictive method and could be replaced by any more appropriately assessed prior. We would, however, stress that the use of the improper prior in these medical diagnostic problems of limited data is a lesser fault than the ignoring of the sampling variability in the estimative approach. When the past experience is large, the difference between the two methods can lessen to an extent which leaves no practical difference; see, for example, the medical examples analysed in Hermans and Habbema (1975). But the question of what constitutes a large set of past records is particularly tricky when multidimensional space is concerned. A set of 100 case records involving 2-dimensional observation vectors is certainly large, but can the same be said for a set of 100 case records involving 85-dimensional vectors? Since the predictive method is just as simple as the estimative method there seems no real need to ask such a question? The answer is indeed simple: be sensible, be realistic, apply the predictive method.

Acknowledgements

The authors are grateful to Dr D. M. Titterington for helpful comments and for making available his tests of multinormality on the data of the illustrative example of this paper; and to the M.R.C. Blood Pressure Unit, Western Infirmary, Glasgow, for making available to us their data. This research was supported for one of the authors (J. W. K.) through a S.R.C. Research Studentship. We also wish to thank the two referees for helpful criticisms.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

Aitchison, J. and Kay, J. W. (1975). Principles, practice and performance in decision making in clinical medicine. *Proceedings of the* 1973 *NATO Conference on the Role and Effectiveness of Theories of Decision in Practice* (D. J. White and K. C. Bowen, eds). London: Hodder and Stoughton.

Aitchison, J. and Sculthorpe, D. (1965). Some problems of statistical prediction. *Biometrika*, **52**, 469–483.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Andrews, D. F., Gnanadesikan, R. and Warner, J. L. (1973). Methods for assessing multivariate normality. In *Multivariate Analysis III* (P. R. Krishnaiah, ed.), pp. 95–116. New York: Academic Press.

Brown, J. J., Fraser, R., Lever, A. F. and Robertson, J. I. S. (1971). Hypertension: a review of selected topics. *Abstracts of World Medicine*, Vol. 45, No. 9.

Dixon, W. J. (1970). *BMD Biomedical Computer Programs*. University of California Press.

Dunsmore, I. R. (1966). A Bayesian approach to classification. *J. R. Statist. Soc.* B, **28**, 568–577.

Geisser, S. (1964). Posterior odds for multivariate normal classifications. *J. R. Statist. Soc.* B, **26**, 69–76.

GUTTMAN, I. and TIAO, G. C. (1964). A Bayesian approach to some best population problems. *Ann. Math. Statist.*, 35, 825–835.

HERMANS, J. and HABBEMA, J. D. F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Med. und Biol.*, 6, 14–19.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press.

PEARSON, K. (1922). *Tables of the Incomplete* Γ*-function.* Cambridge University Press.

—— (1934). *Tables of the Incomplete B-function.* Cambridge University Press.

TAYLOR, T. D., AITCHISON, J. and MCGIRR, E. M. (1971). Doctors as decision makers: a computer-assisted study of diagnosis as a cognitive skill. *Br. Med. J.*, 3, 35–40.

ZELLNER, A. and CHETTY, V. K. (1965). Prediction and decision problems in regression models from the Bayesian point of view. *J. Amer. Statist. Ass.*, 60, 608–616.

11 THE ANALYSIS OF SUBJECTIVE PERFORMANCE IN INFERENTIAL TASKS

Considerable light can be thrown on the intuitive or subjective processes by which individuals arrive at inferences. For example, it is possible to examine in some detail how clinicians, faced with the problem of making diagnostic assessments, differ in their assessments from those reached by a statistical diagnostic system such as predictive diagnosis. Such an analysis can serve two purposes: first, to reveal where the intuitive, subjective judgment may be falling short of the optimum or normative; secondly, by showing the clinician ways in which he may be misusing the information available to him, to exploit the neat expository aspects of the comparison to engage his interest in learning more of inferential principles and methodology. Such studies have been reported in the differential diagnostic problems of non-toxic goitre (Taylor, Aitchison and McGirr, 1971), and in a simulated diagnostic situation in the differential diagnosis of newmath fever in students (Aitchison and Kay, 1973; Aitchison, 1974), with illustrations of how the performance may be analysed and presented, possibly sequentially as each new feature of the patient is disclosed. Aitchison and Kay (15:1975) and Kay (1976) give details of the various measures of performance which throw light on the discrepancies between subjective and 'normative' performance. Where past experience is controlled by the presentation only of a training set then normative is taken to be equivalent to the predictive diagnostic assessment.

If the set of possible diagnostic types is T then a

diagnostic assessment about type is simply a probability density function p(t) on T. For any such assessment there is a *degree of uncertainty* remaining, namely

$$U\{p(t)\} = -\Sigma_T\, p(t) \log p(t).$$

If from an established position p(t) a subject, on the basis of evidence x, moves to s(t|x) instead of to r(t|x) then his *inference discrepancy*, a measure of his inability to make the correct inference, is

$$I\{r,s\} = \Sigma_T\, r(t|x) \log \frac{r(t|x)}{s(t|x)}.$$

Moreover in his move from p(t) to s(t|x) the subject removes an amount of uncertainty U{p(t)} - U{s(t|x)} instead of the appropriate amount U{p(t)} - U{r(t|x)}. The difference in these

$$|\, U\{r(t|x)\} - U\{s(t|x)\}\, |,$$

suitably signed, can then be used as a measure of whether the subject is underusing the data (negative sign) or reading too much into the data (positive sign). For more details of this *information gain index* (positive or negative) see Aitchison and Kay (15:1975).

Part of the diagnostic task may be to choose, after each assessment, the next feature to be observed or the next test to be carried out from a still available set F of such features or tests. The gain of information from the choice of $f \in F$ and observing x, and hence updating from p(t) to $r_f(t|x)$ is $U\{p(t)\} - U\{r_f(t|x)\}$. At the time of choosing such a test the expected gain of information is

$$G(f) = \int_X [U\{p(t)\} - U\{r_f(t|x)\}]\, p_f(x)\, dx.$$

Let f* be a feature maximising G(f). Then if the subject chooses f his feature *selection discrepancy* is measured as

$$G(f^*) - G(f).$$

Thus there is a whole battery of measures which display different aspects of inference ability in relation to the quantified inference through the use of parametric model fitting. Experience suggests that these performance analyses are excellent means of motivating teaching and of exposition. There is no real evidence that subjects improve their subjective ability in making inferences by exposure to them.

The form of performance analysis which has been illustrated here in a diagnostic setting applies equally for any other inferential task where the subject is required to specify the equivalent of a probability density function over a set T. For example, Aitchison (1980b) describes a calibration task. Similar concepts apply in the analysis of performance in decision-making tasks; see, for example, Aitchison et al (1973) and Aitchison and Moore (1976).

A more recent development, which allows investigation of the detailed inferential statements rather than summary measures of them, is reported in §13.

AITCHISON, J. and KAY, J.W. (1975)

Principles, practice and performance in decision-making in clinical medicine

Reprinted from *The Role and Effectivenss of Decision Theories in Practice* (eds K.C. Bown and D.J. White): London: Hodder and Stoughton, pp.252-72

Principles, Practice and Performance in Decision Making in Clinical Medicine

J. AITCHISON

Professor of Statistics, and

J. W. KAY

Department of Statistics,
University of Glasgow, U.K.

1. INTRODUCTION

Few people face, and have to resolve, a more intensive stream of important decision problems during a working lifetime than the clinician; and few are such inveterate collectors of data and such meticulous recorders of case histories. Clinical medicine therefore seems a promising area of human activity in which to study the role and effectiveness of decision theories in practice. Moreover, it is an area which differs in many respects from the more common target of decision theorists—the business and industrial world. For example, while, in business, an acceptable aim for a number of enterprises may be to maximise average profit over these enterprises, in clinical medicine, the Hippocratic oath precludes any criterion of average results over individual patients. Any sensible applied mathematician—and we hope that most decision theorists fall into this category—does not start with a theory and seek areas of application for that theory. Rather he starts with a real problem, studies it until he feels able to abstract the essential relevant components, to recognise their interdependence, to express this in the language of mathematics and so build a mathematical model; he then develops his model mathematically in a sensible direction towards the resolution of the problem initially posed; and finally he translates his mathematical answer back into real terms and so may be in a position to assess the effectiveness of his model-building. In comparing clinical medicine and business, therefore, we must not be surprised if we find differences in the decision model itself, in the emphasis of its component parts and in the mode of application.

Three basic questions immediately arise. Can we find in clinical medicine any acceptable principles of action which may be formalised and translated into decision-theoretic terms? To what extent can we apply any such model of the clinical decision process in practice? In what meaningful ways can we compare the performance of the decision model with that of the clinician? While it is easy to speak speculatively of a decision system which embraces the whole of medical

practice, its implementation is no more than a current pipedream. To attempt to assess role and effectiveness, we must clearly limit ourselves to such sub-systems as are currently practicable. To accept such a delimitation lays us immediately open to the critical question: could you not arrive at better decisions by embedding the subsystem in a larger system? While this implied criticism is strictly unanswerable, we hope that the evidence we shall put forward of the remarkable discrepancies in practice and performance between different decision subsystems may persuade the reader that their study must shed some light on optimal decision-making.

There are, in clinical medicine, two main ways in which subsystems traditionally arise. The first is the streaming of patients into specialities according to their suspected condition; all the specialist clinics, such as blood pressure, thyroid, renal, psychiatric, in our hospitals are a result of this partition. The second is the division of the management of a patient into traditional phases—examination, diagnosis, treatment, prognosis, after-care. While we shall recognise these two forms of subdivision, we shall not adopt them uncritically. We shall keep watch on the first by a monitoring of each patient to check on the appropriateness of his streaming. We shall question the second by considering the synthesis of the traditional subdivisions and re-examining the appropriateness of these subdivisions within the synthesis.

We must also recognise that, at any moment of time, we can discern two aspects within the medical decision maker—the doctor whose sole concern is for the patient he is currently managing, and the scientist who hopes to obtain from his patient some deeper insight into medical science, hopefully to the advantage of subsequent patients. In the event of a clash of objective between these two roles, the Hippocratic oath dictates that the role of scientist must be subordinate to the role of doctor.

In outline, the development proceeds as follows. In section 2 we sketch the main components of clinical decision, making no attempt at precision but rather painting a broad picture of the subject and establishing some terminology and notation. A central problem in current medical practice is diagnosis, and we examine in detail the construction of a realistic diagnostic model—the predictive diagnostic model—in section 3, compare its effectiveness in practice with widely advocated alternatives in section 4, and discuss its use as a basis or norm for measuring diagnostic performance in section 5. The closely related aspects of prognosis, treatment assessment and treatment allocation are examined in section 6 and a synthesis of the various phases of patient management considered in section 7.

An impetus to the re-examination of clinical medicine in decision theoretic terms was given by Ledley and Lusted[1]; see also Lusted[2] for a more extensive account and some interesting applications. We shall not attempt to provide a complete survey of the subject here or to trace the source of all the ideas presented. Rather we shall concentrate on problems which have come our way in the course of providing a consultative service to medical colleagues, and try to draw some conclusions from this experience.

2. *COMPONENTS OF CLINICAL DECISION MAKING*

When faced with any problem man seems to find it an aid, or at least a consolation, to be able to give it a label. The medical profession is no exception; its concept of diseases is essentially one of classification. The clinician is regarded as facing a stream of patients each belonging to one and only one of a finite set *C*

$= \{1, \ldots, d\}$ of *disease categories.* The 'only one' provision can be ensured by having separate categories for each feasible multiple pathology. Much energy is indeed expended in the formulation of the appropriate C for various specialties. A presenting patient is of unknown category κ although the clinician, from his experience may have formed some view of the arrival pattern. This we can characterise by the *arrival rate* vector α of probabilities α_c, where α_c is the probability that the next patient will be in disease category c. The clinician will not normally know α but may regard some α as more plausible than others. Conceptually any α in the simplex

$$A = \{\alpha : \alpha_c > 0 \, (c = 1, \ldots, d), \alpha_1 + \ldots + \alpha_d = 1\} \tag{2.1}$$

is feasible. We have placed the strict inequality on each α_c to ensure that no category can ever be ruled out with absolute certainty.

In the diagnostic phase of patient management the clinician gathers information about some or all of a finite set $F = \{1, \ldots, g\}$ of *features.* Feature 1 might be sex, feature 2 presence or absence of headache, feature 3 pulse rate, feature 4 I^{131} 24-hr uptake, and so on; in other words, observed features may be personal information and symptoms elicited from the patients, signs observed or measurements made by the clinician at his examination of the patient, results of laboratory tests, etc. For this preliminary view of the process, we shall not dwell on the problem of which order and which stopping rule should be used in this accumulation of information. We shall assume that the whole of F has been observed, and that the appropriate sample space is X_F, or more briefly X. We shall use suitably indexed x's to denote the observed feature vectors of patients of known category; for the current undiagnosed patient, we shall denote the observed feature vector by y.

If the categories are really distinct and the features are to be of any value in the direction of category, we must envisage that the distribution of x depends to some extent on the category c of the patient. We can exhibit this dependence by writing $p(x|c,\theta_c)$ for the probability (density) function associated with category c. The θ_c is also introduced to emphasise that we do not know what this distribution is, but have to imagine it as one of some class of possible distributions indexed by elements of Θ_c. We lose no generality, and we shall obtain some notational simplicity, if we write $\theta = (\theta_1, \ldots, \theta_d)$ for the *structural parameter,* and $\Theta = \Theta_1 \times \ldots \times \Theta_d$ and consider the feature vector distribution associated with category c as belonging to a class of distributions

$$\{p(\cdot|c,\theta) : \theta \in \Theta\} \tag{2.2}$$

on X.

For convenience we use the term *case record* for the combination (c,x) associated with a patient whose feature vector is x and who has been classified firmly into category c. We suppose that there is available a number n of such case records, n_c of category c, where $n_c > 0$ for every $c \in C$. We denote this set of *past records* briefly by

$$z = \{(c_i, x_i) \; : \; i = 1, \ldots, n\} \tag{2.3}$$

Within this framework diagnosis is the process by which the clinician uses whatever information he has concerning the unknown α and θ, together with past records z and the data y on his new patient, to arrive at an assessment of the plausibilities

$$p(\kappa|y,z;h) \tag{2.4}$$

of the possible categories $\kappa = 1,\ldots,d$ for the new patient. We deliberately introduce an h into the notation to emphasise that this assessment must depend on certain assumptions or hypotheses and it is well to be reminded forcefully of their presence. We shall consider in detail in section 3 the diagnostic aspects of the process and, in particular, the hypotheses h underlying the process.

Having hopefully reduced, in some sense, the dimensionality of the problem by this technique, the simple traditional view is to turn attention to the problem of allocating an appropriate treatment. While there may be situations in medicine where there is one-to-one correspondence between category and treatment, there are usually options open at the end of the so-called diagnostic process. We therefore recognise that there may be a set T of possible *treatments,* and since we regard diagnosis as a probability assessment rather than a decision on category, we assume that T contains any known treatment appropriate to any of the categories in C. We also recognise that the response to the same treatment t of two patients, each with the same (c,x), will not necessarily be the same. Let us suppose that assessment of the effectiveness of a treatment is measured in terms of observations of a finite set $R = \{1,\ldots,s\}$ of *responses* or features, some of which may, of course, be the same as prior to treatment. It may also be that, between diagnosis and treatment, information additional to the diagnostic y information is required. We can, however, assume that this is absorbed in the y displayed. Suppose that the sample space associated with observation of R is W_R, or simply W.

In his search for an appropriate treatment for a patient with known case record (c,x), the clinician prognoses: if he assigns treatment t what is the response w likely to be, or more appropriately what is the probability density function $(w|c,x,t)$ on W? In order to compare the effectiveness of different treatments, it is necessary to have some knowledge of this. To emphasise the fact that we may not know this exactly, we write it as

$$p(w|c,x,t,\psi) \tag{2.5}$$

to indicate that we are dealing with one member of a family of possible distributions indexed by $\psi \in \Psi$. The response w may in some cases be measured in simple terms such as cure or no cure, or survival time, but there are clearly situations where some much more complex measurement is more realistic.

The distributions (2.5) describe the prognostic aspect of medicine and are conveniently called the *prognostic distributions.* The source of any firm information on them or equivalently on ψ is undoubtedly the controlled clinical trial; and from such information, we would attempt to eliminate the nuisance index ψ to obtain applicable prognostic distributions

$$p(w|c,x,t;h) \tag{2.6}$$

for each possible c,x,t. Again we insert a cautionary h to remind us of the assumptions involved.

To allocate treatment t to a new patient (with feature vector y but of unknown category κ), we have to have some overall view of the advantages and disadvantages of the various possible outcomes. If the clinician could make explicit a general utility structure

$$U(\kappa,y,t,w), \tag{2.7}$$

the utility of getting a patient with κ from unpleasant state y by the discomfort and cost of treatment t to the more pleasant state w, then standard approaches to decision demand that we choose t to minimise

$$\sum_{w \in W} \sum_{\kappa \in C} U(\kappa,y,t,w)\, p(\kappa \mid y,z,;h)\, p(w \mid \kappa,y,t;h). \tag{2.8}$$

3. *A DIAGNOSTIC MODEL*

We now attempt to specify clearly the basic assumptions of statistical diagnosis. We have tried to reduce the assumptions to the absolute minimum and to express them in a form in which they have as far as possible a practical meaning. Certainly only by setting out precisely what the assumptions are is it possible to discuss the appropriateness of analysis that depends on these assumptions. It will be seen that with five such assumptions, h1 to h5, some rather interesting consequences follow. The notation follows that of section 2. For a full discussion of the appropriateness of these assumptions see Aitchison and Kay.[3]

h1: $p(c \mid \alpha,\theta) = p(c \mid \alpha)$.

h2: $p(x \mid c,\alpha,\theta) = p(x \mid c,\theta)$.

h3: For *any* set z of m case records z_i $(i = 1, \ldots, m)$

$$p(z \mid \alpha,\theta) = \prod_{i=1}^{m} p(z_i \mid \alpha,\theta).$$

h4: $p(\alpha,\theta) = p(\alpha)p(\theta)$.

We do not imply in the notation $p(c \mid \alpha)$ that the cases considered as constituting past records have necessarily arisen at the natural arrival rate. We allow the possibility for instance that the informative experiment has been specially designed so that there are equal numbers of cases in each of the categories. The new case for consideration, of unknown category κ, is however assumd to arrive according to the probabilistic pattern associated with α.

h5: $p(\kappa \mid \alpha) = \alpha_\kappa$, where α_κ is the κ element of α.

These assumptions, together with straightforward conditional probability arguments, including Bayes' Theorem, allow us to obtain the form of the posterior plausibility assessment α, θ, κ, based on the information y, z; for details see Aitchison and Kay.[3] Thus,

$$p(\alpha,\theta,\kappa \mid y,z;h) = \frac{p(\alpha \mid c)p(\kappa \mid \alpha)p(\theta \mid z)p(y \mid \kappa,\theta)}{\sum_{\kappa \in C} p(\kappa \mid c)p(y \mid \kappa,z)} \tag{3.1}$$

where $c = (c_1, \ldots, c_n)$, $x = (x_1, \ldots, x_n)$, and

$$p(\alpha|c) = p(\alpha)p(c|\alpha)/\int_A p(\alpha)p(c|\alpha)d\alpha, \tag{3.2}$$

$$p(\theta|z) = p(\theta)p(x|c,\theta)/\int_\Theta p(\theta)p(x|c,\theta)d\theta, \tag{3.3}$$

$$p(\kappa|c) = \int_A p(\kappa|\alpha)p(\alpha|c)d\alpha, \tag{3.4}$$

$$p(y|\kappa,z) = \int_\Theta p(y|\kappa,\theta)p(\theta|z)d\theta. \tag{3.5}$$

Although we shall be interested in other aspects of this posterior distribution, our first objective is to assess the different plausibilities of the various categories, that is

$$p(\kappa|y,c,x;h) = \frac{p(\kappa|c)p(y|\kappa,c,x)}{\sum_{\kappa\in C} p(\kappa|c)p(y|\kappa,c,x)} \tag{3.6}$$

An interpretation of (3.6) is that it is simply the conversion of an assessment $p(\kappa|c)$, after the categories c of past cases are known, but prior to any information concerning the $n+1$ feature vectors y,z, to a posterior assessment $p(\kappa|y,c,x;h)$ by way of Bayes' Theorem and with $p(y|\kappa,c,x)$ or $p(y|\kappa,z)$ playing the role of the likelihood function. Special interest then centres on $p(\kappa|c)$ and $p(y|\kappa,x)$.

First we note that, from (3.4) and h5,

$$p(\kappa|c) = E(\alpha_\kappa|c). \tag{3.7}$$

Hence in so far as inference concerning the category of the new case is concerned, uncertainty about α is involved only in the form $E(\alpha_\kappa|c)$. We do not have to be able to provide a complete picture of the uncertainty in $p(\alpha|c)$ but only the mean vector $E(\alpha|c)$ of this distribution.

The distribution $p(y|\kappa,z)$ defined by (3.5) is the *predictive distribution.* For a new case in *known* category κ it provides, on the basis of prior information $p(\theta)$, the past records z and the assumptions h, an assessment of the probabilities of the possible feature vectors y we may observe on the case. The important role that predictive distributions have to play in statistical analysis has become increasingly recognised (Geisser[4,5], Guttman and Tiao,[6] Aitchison and Sculthorpe,[7] Zellner and Chetty,[8] Dunsmore,[9,10,11] Lindley[12]), and it is not new to the field of statistical diagnosis. The relation (3.6) is the basis of a method advocated by Geisser[4] in relation to multivariate normal data and by Dunsmore[9] in a decision theory approach to the general problem of identification. Neither gives applications to clinical medicine. As far as we are aware no such practical application has been made nor have the practical implications of the method been followed through. We have redeveloped the model so as to emphasise, and give meaning to, the underlying assumptions, to allow consideration in later sections of the full posterior dependence of α, θ and κ to obtain some wider results, and to set the scene for the assessment of its effectiveness in practice and for the analysis of performance.

Although the diagnostic phase is a means to an end it is undoubtedly an important phase, and it is possible at this stage to make some assessment of the model concerned, both from the practical viewpoint and from the performance. Such assessments are the subjects of the next two sections.

4. THE DIAGNOSTIC MODEL IN PRACTICE

A few simple assumptions have led us inevitably to a particular form of diagnostic assessment, which we shall term *predictive* diagnosis because of its association with the predictive distribution. We now compare the practical effectiveness of this method with others commonly used.

The statistical diagnostic methods currently widely advocated (for example, Dixon[13]) are of *estimative* type and are generally supported by an argument of the following kind. If we knew the structural parameter θ then we could easily arrive at appropriate plausibilities for the unknown disease category κ of a new patient with feature vector y. We would simply apply Bayes' Theorem in the form

$$p(\kappa|y) \propto p(\kappa)p(y|\kappa,\theta). \tag{4.1}$$

Recognising that we do not know θ, the estimative method simply replaces θ by a suitable estimate $\hat{\theta}(z)$, for example a maximum likelihood estimate, based on the past records z. We could thus rewrite the estimative method as

$$p(\kappa|y,z) \propto p(\kappa|c)p(y|\kappa,\hat{\theta}(z)). \tag{4.2}$$

Using a simpler form of (3.6) for the predictive method, and for brevity dropping the h,

$$p(\kappa|y,z) \propto p(\kappa|c)p(y|\kappa,z), \tag{4.3}$$

where

$$p(y|\kappa,z) = \int_{\Theta} p(y|\kappa,\theta)p(\theta|z)\mathrm{d}\theta. \tag{4.4}$$

That the difference between the two techniques can have a substantial effect in practice can be seen from the following example of the application of the two methods. Until fairly recently, a rare hypertensive syndrome (Conn's syndrome) was believed to have as its sole possible cause an adenoma (benign tumour) in an adrenal cortex. At operation on a number of patients, several were found not to have an adrenocortical tumour but to have a more diffuse condition (bilateral hyperplasia) involving both adrenal glands. Since the assessment of treatment, which may range from total adrenalectomy, through removal of an adenomatous adrenal gland if locatable, to drug therapy, is now recognised to depend on the diagnostic assessment and on a number of factors external to the diagnostic assessment, what is required from the diagnostic process is a reasonable assessment of odds. There are eight features or diagnostic tests and the basic set z of past records consists of 31 case records, 20 in category a (adenoma) and 11 in category b (bilateral hyperplasia). There is appreciable evidence of non-equality of covariance matrices, so that the estimative method would be attempting to estimate an 88-dimensional parameter from 31 eight-dimensional vectors. It is clear that to use these estimates as if they were reliable is running a risk. A first indication of the remarkable extent of this risk is seen in figure 1, where the $\log_{10}$ odds (a/b) based on the estimative and the predictive methods are shown for the 31 patients of the basic set. Odds of 10^{20} to 1 by the estimative assessment are slashed to 10^3 to 1 by the predictive assessment. The position with respect to new patients can be even more startling. Table 1 shows experience with new patients. The dramatic alteration

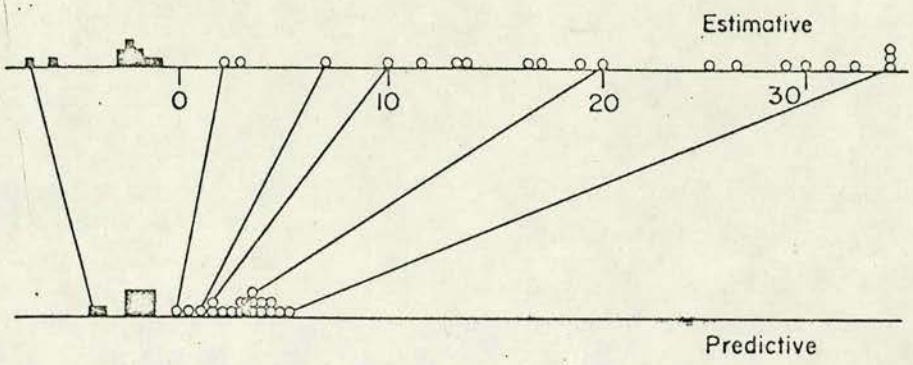

FIGURE 1 Comparison of the $\log_{10}$ odds (a/b) assigned by estimative and predictive diagnosis to the 31 basic case records

of odds in some of the cases is due to the fact that they are, in a technical sense which can be well defined, outside the previous limited experience of both diseases, though not in any way atypical of the favoured disease. For such cases, the predictive method acts with more caution, which is again very reasonable. Cases N4 and N5 are particularly interesting. Case N4 is in fact a now-confirmed case of bilateral hyperplasia, not the clear case of adenoma assessed by the estimative method. The true category of case N5 is as yet unknown but treatment on the basis of a diagnosis of bilateral hyperplasia is so far meeting with success.

TABLE 1 Comparison of estimate and predictive odds for 5 new patients with Conn's syndrome

Patient Number	Odds a/b	
	Estimative	Predictive
N1	$10^{12}/1$	64/1
N2	$10^{18}/1$	56/1
N3	1/245	1/80
N4	3700/1	1/2
N5	$10^{9}/1$	1/3

In our introductory remarks, we indicated that some form of monitoring is desirable to ensure that the decision to stream a patient into the specialty characterised by C has not been unreasonable. A method of achieving this is based on the following argument. On the basis of past experience of disease category the probability distribution associated with the feature vector y of any new patient is simply the predictive distribution. Suppose we are concerned about how typical of disease category κ a patient with observed feature vector y_0 is? We can regard any case with feature vector y for which $p(y|\kappa,z) \geqq p(y_0|\kappa,z)$ as more typical of κ than patient y_0, or equivalently y_0 as less typical than y. We can thus construct for patient y_0 a sensible index $A_\kappa(y_0)$ of atypicality relative to disease category κ as

the probability (on the basis of past experience) that another patient is more typical than him. Thus

$$A_\kappa(y_0) = \int_{\{y:\, p(y|\kappa,z) \geqslant p(y_0|\kappa,z)\}} p(y|\kappa,z)\,dy.$$

Thus A_κ is measured on the scale (0,1) with 0 indicating the absolutely typical and 1 complete atypicality. If we find

$$\min_{\kappa \in C} A_\kappa(y_0)$$

near 1, then we would be right to suspect that the patient may have been channelled into the wrong set C of categories, and take some appropriate action.

5. *MEASURING PERFORMANCE IN DIAGNOSTIC TASKS*

We hope that by now we may have convinced the reader of the reasonableness of predictive diagnosis as a means of assessing the diagnostic position of a particular new patient. We propose shortly to use predictive diagnosis as a normative model against which to assess the intuitive performance of human decision makers in such diagnostic tasks. Even for the reader who does not accept the predictive diagnostic model, it is easy to demonstrate the great variability of inferential ability and how it falls short of generally accepted inference. We can achieve this by ensuring that the decision maker knows the distributions associated with the three categories.

TABLE 2 Composition of the three possible boxes *a, b, c*. The numbers shown are the numbers of black beads in each compartment

Box	Compartment 1	2	3	4	5	6	7	8	9	10
a	4	8	7	4	6	5	3	2	4	1
b	4	7	4	8	5	9	5	3	8	5
c	6	6	6	5	2	4	8	7	2	9

A simple form of diagnostic task is in terms of a set of independent diagnostic binomial tests which can be described in terms of a box model. Suppose that a box is of one of three possible types *a*, *b* or *c*. Each box is divided into 10 compartments, each compartment containing 10 beads some black some white. Table 2 shows the numbers of black beads in each of the compartments in each of the boxes, and this information is supplied to the subject or decision maker. The subject is then informed of the results of a single random drawing from each of the 10 compartments of a box of undisclosed type; the results are given in sequence and he is asked to update his plausibility assessment after each result is disclosed to him. A convenient way to express the opinion that the relative plausibilities of *a*, *b* and *c* are π_a, π_b, π_c, with $\pi_a + \pi_b + \pi_c = 1$, is in terms of the point π in the equilateral triangle abc of unit altitude (figure 2). Roughly speaking with this representation the nearer a point is to a vertex the more plausible the corresponding

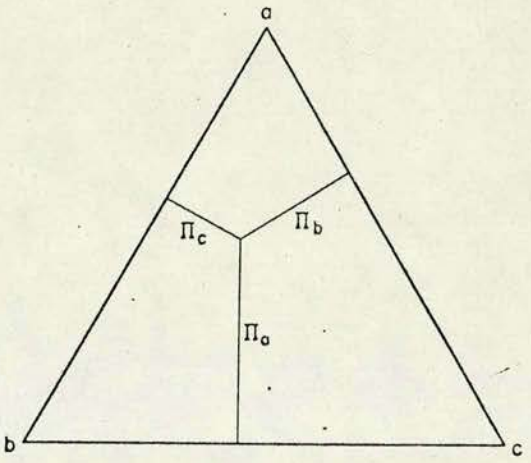

FIGURE 2 Representation of a plausibility assessment within a triangle

box or disease category is regarded. Thus for each case presented to him the subject will produce a diagnostic path of 10 steps showing the development of his plausibility assessment as the evidence accumulates. Figure 3 shows diagnostic paths for a variety of individuals for three cases. It is clear that there is wide variation in ability and departure from the objective path computed by repeated application of Bayes's Theorem.

For the case where the feature distributions are not known but have to be visualised from case records by the subject, the variability of diagnostic inference can again easily be demonstrated. At a recent conference on multivariate statistical analysis a diagnostic competition was held. The observed vectors on a six-dimensional feature on 48 patients, 16 in each of three disease categories *a,b,c,* were given to each contestant, with, for each category sample, the feature means, standard deviations and correlation coefficients already computed. The contestants were then asked to provide a diagnostic opinion on each of five uncategorised patients for whom the complete feature vectors were provided. They were informed that the diseases were equally prevalent. Figure 4 shows the plausibility points of the 55 participants for one of the cases, a picture which is typical of the other four cases.

For the illustrative case of three diseases we have so far envisaged the decision maker as making his diagnostic assessments on a flat triangular plane. We can, however, give an added dimension to the picture and an added depth to our understanding of this phase of the decision making process by the introduction of information concepts. We shall do this through the familiar concept of degree of uncertainty (the so-called entropy of information theory) remaining about the unknown category κ, when the plausibility assessment for κ is π_κ ($\kappa \in C$). This is defined by

$$u(\pi) = -\sum_{\kappa \in C} \pi_\kappa \log \pi_\kappa, \qquad (5.1)$$

where the base of the logarithmic function determines the units. The greatest degree of uncertainty occurs in the equi-plausible case $\pi_a = \pi_b = \pi_c = \frac{1}{3}$, when $u(\pi) = \log 3$; and the least approaches 0 when π_κ approaches 1 for some $\kappa \in C$.

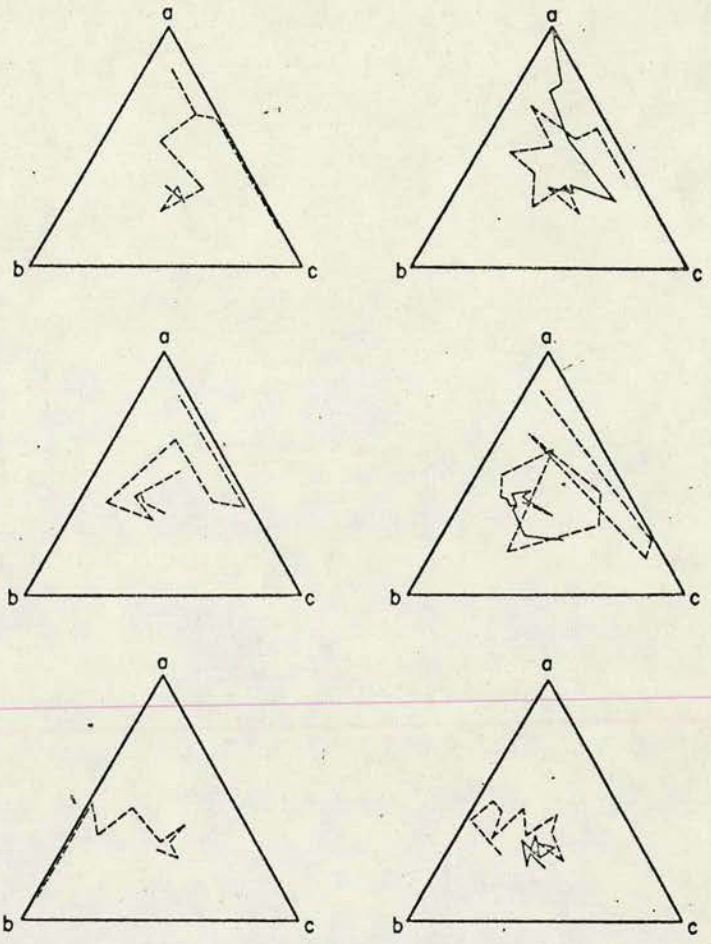

FIGURE 3 Diagnostic paths for three cases of trilemma. The objective paths are shown in the triangles on the left. The subjects whose paths are shown on the right are:
(i) Two consultant clinicians
(ii) Two first-course statistics students
(iii) ——— 10 year old boy
— - - - - - - consultant clinician

Each point in the triangle has therefore associated with it a degree of uncertainty and we can draw within the triangle the contours of uncertainty consisting of points (π_1, π_2, π_3) for which $u(\pi)$ is constant (figure 5). If we take this extra dimension 'degree of uncertainty' as depth below the surface of the triangle, we obtain a picturesque view of the decision maker in his search for the unknown category. He is in a pit of uncertainty shaped like a triangular bowl, endeavouring, by his feature observations, to gain enough information so that he may climb sufficiently near the vertex corresponding to the true category and thus make the correct diagnosis. The concepts presented here for the case $d = 3$ extend without modification for other d, although of course the visual representation is lost.

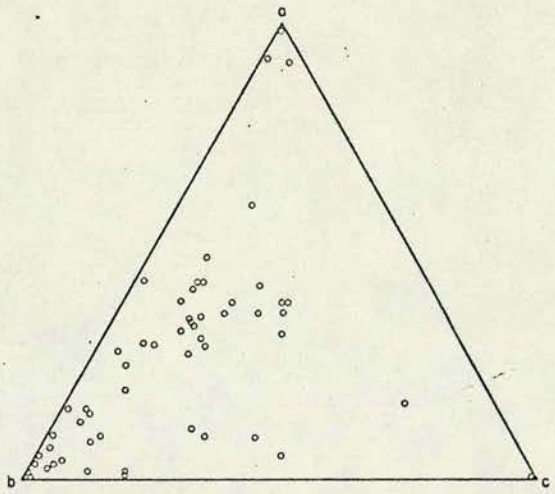

FIGURE 4 Plausibility points of 55 participants in a diagnostic competition

The question of how to obtain meaningful measures of the discrepancy between the personal inference of the diagnostician and the predictive diagnosis assessment has now to be tackled. We shall take as our general objective the reduction of the degree of uncertainty about the unknown category and we shall judge a decision maker's effectiveness in terms of his ability to select his features for this purpose, and also his skill in interpreting the feature observations as indicators of the true category.

Suppose that at any stage in his diagnostic path the decision maker's assessment is represented by plausibilities π_κ, $\kappa \in C$, and that after observing an additional feature, he quotes π'_κ, instead of p_κ given by the predictive diagnostic model, as his new assessment (figure 6).

The first measure of discrepancy is a general one which is positive except when $\pi'_\kappa = p_\kappa$ for every $\kappa \in C$, when it assumes the value zero. We shall term it the *inference discrepancy* $\mathcal{I}$, defined by

$$\mathcal{I}(\pi, \pi', p) = \sum_{\kappa \in C} p_\kappa \log (p_\kappa / \pi'_\kappa) \quad . \tag{5.2}$$

This measure may be computed for each step in the decision maker's diagnostic path. At the beginning of each step, we place the decision maker in the most favourable position by accepting, in the computations for that step, his attained plausibility assessment. Each measurement $\mathcal{I}$ is therefore made relative to the particular step and does not contain any accumulation of discrepancies from previous steps.

We now introduce measures of discrepancy for conservatism and liberalism. If the decision maker moves from π to π', and π' is below p on the uncertainty surface, then we can meaningfully say that he is acting in a conservative manner relative to the normative model. The extent of this conservatism can be measured by the difference in the uncertainty levels of p and π', namely

$$u(p) - u(\pi'),$$

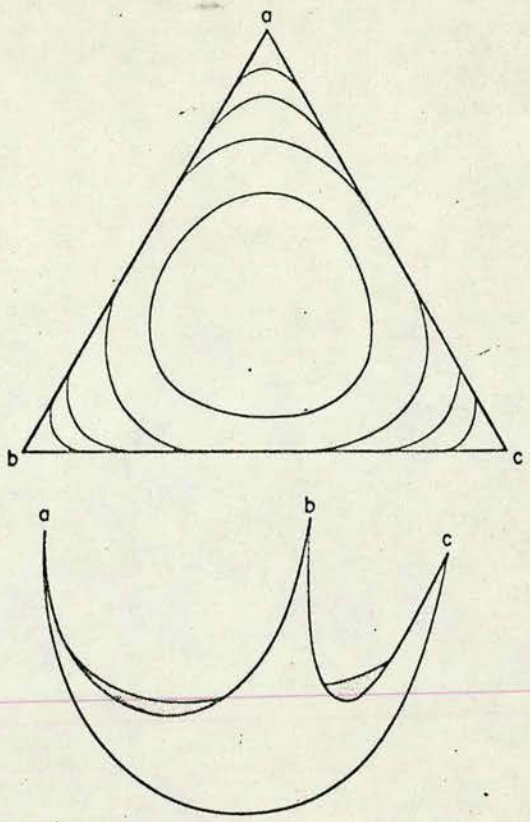

FIGURE 5 Contours of uncertainty and triangular bowl

where the difference has been expressed so that a conservative use of a feature observation is associated with a negative measure. If π' is above p on the uncertainty surface, the decision maker is reading more into the feature observation than the normative model allows. This liberalism in the interpretation of the feature observation can be measured by $u(\pi') - u(p)$, the fact that this is positive indicating liberalism and not conservatism.

The construction of this conservative-liberal index $\mathcal{L}$ may be expressed as

$$\begin{aligned} \mathcal{L}(\pi, \pi', p) &= u(p) - u(\pi'), \text{ if } u(\pi) > u(p), \\ &= u(\pi') - u(p), \text{ if } u(\pi) < u(p), \end{aligned}$$

and, if $u(\pi) = u(p)$, (5.3)

$$\begin{aligned} \mathcal{L}(\pi, \pi', p) &= u(p) - u(\pi'), \text{ if } u(\pi) \geqslant u(\pi'), \\ &= u(\pi') - u(p), \text{ if } u(\pi) < u(\pi'). \end{aligned}$$

Suppose that we take as our criterion of feature selection that of expected gain of information, or equivalently expected reduction of information from observing a feature or subset of features. Following Lindley,[14] we may then define $G(J|y_I)$, the

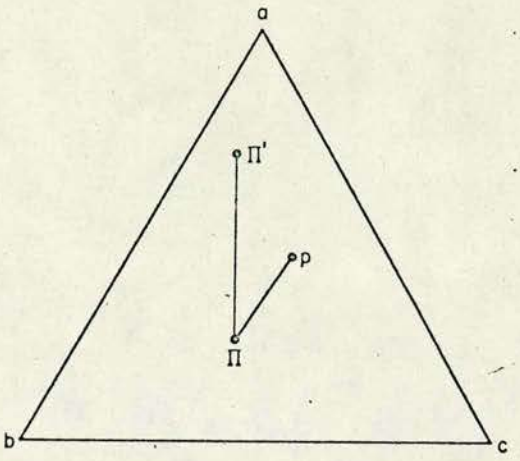

FIGURE 6 Comparative steps in a diagnostic path

expected gain of information about κ from observing the subset J of features after y_I has already been observed in the subset I of features. In the normative model, we decide to observe the set of features J^*, which has maximum expected gain of information over all subsets J of F which have no features in common with I. Relative to this normative model, a measure $\mathscr{F}$ of feature selection discrepancy for a subject who chooses J after observing y_I is given by

$$\mathscr{F}(y_I) = G(J^*|y_I) - G(J|y_I).$$

This quantity may be computed for each step ($j = 1$) in the diagnostic path.

To illustrate the method of analysing performance we show in figure 7 the sequential diagnostic paths and the associated performance analyses of two subjects for a new case of the three-disease system referred to at the beginning of this section. Taylor *et al.*[15] give an application in a medical setting and show similar discrepancies for clinicians, although they are self-critical of their assumption that for a given disease category features are statistically independent. No such assumption is made in the above analysis.

6. *PROGNOSIS AND TREATMENT*

The role of diagnosis in clinical medicine can be regarded as a preliminary phase in which an attempt is made to discover the category or type of the subsequent decision problem of patient management that next faces the clinician. Our emphasis on this phase has been conditioned not only by its obvious importance in current medical thinking but also because it is at present the best quantified phase of most medical problems. Let us now turn our attention to the complex of less-well quantified concepts and actions which are usually considered under the headings of prognosis and treatment.

The main objective tool by which clinicians have attempted to compare and assess treatments is undoubtedly the *controlled clinical trial.* It is beyond the scope of this paper to go into a detailed appraisal of the development and organisation of such trials, the question of ethics and the relative effectiveness of sequential and

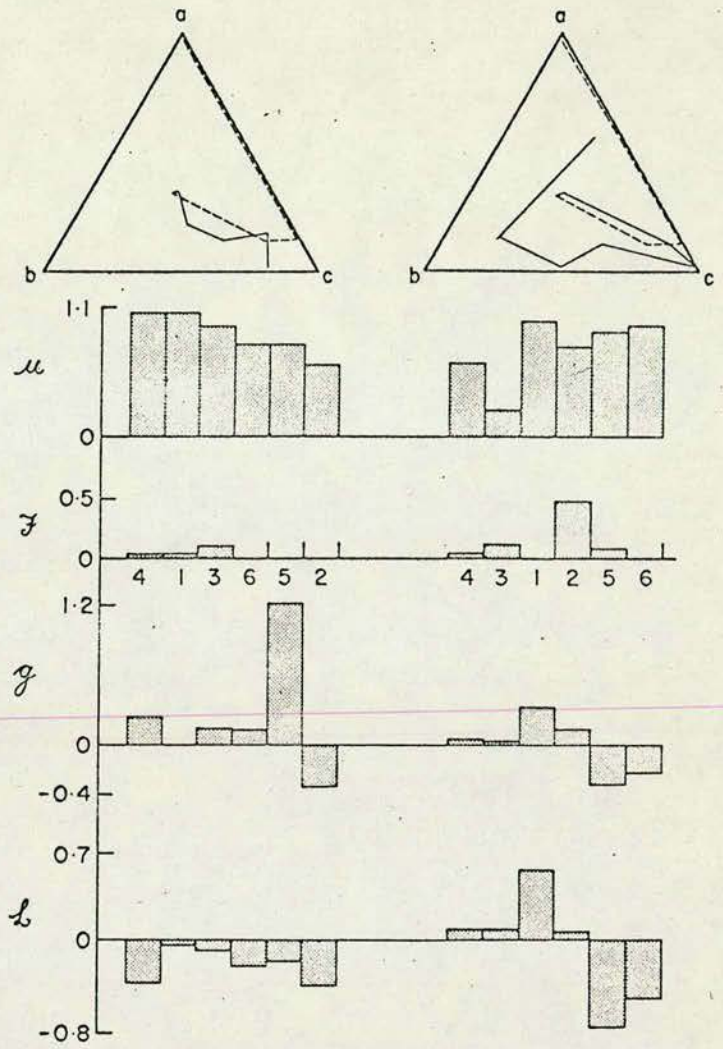

FIGURE 7

fixed-size trials. One general point concerning controlled clinical trials does, however, provide the motivation for our subsequent formulation of the decision problem, and we can here illustrate it by a simple example. Suppose that, in a clinical trial to compare two treatments t_1 and t_2, 200 patients are allocated randomly, 100 to each treatment; that a check is made of the similarity of the composition of the two groups; and that the usual double blind requirements of management of patients and assessment of treatment are met. Suppose that the results of the trial are

	t_1	t_2
Success	50	70
Failure	50	30

The standard statistical analysis would then test the null hypothesis of no difference between the treatments by a standard chi-squared test. A significant difference between the treatments would thus be revealed and, if the sole criterion is to maximise the proportion of successful treatments, the accompanying recommendation would be that treatment t_2 should be used.

If we have to choose between using treatment t_1 *for all patients* and using treatment t_2 *for all patients*, we may be convinced that treatment t_2 is to be preferred. But we may not have made fully effective use of all the information available. For consider the conceptual classification of patients into four mutually exclusive groups e_{11}, e_{10}, e_{01}, e_{00}, where the suffices i, j are assigned for each patient by the following criterion:

$$\begin{aligned} i &= \begin{cases} 1 \\ 0 \end{cases}, \text{if treatment } t_1 \text{ would/would not be successful with the patient,} \\ j &= \begin{cases} 1 \\ 0 \end{cases}, \text{if treatment } t_2 \text{ would/would not be successful with the patient.} \end{aligned} \tag{6.1}$$

It is of course impossible to assign a patient to these groups, but the concept allows us to make the following points. If $p(e_{ij})$ denotes the proportion of patients in e_{ij} in the trial, then the only restrictions determined by the results are

$$\begin{aligned} p(e_{11}) + p(e_{10}) &= 0.5, \\ p(e_{11}) + p(e_{01}) &= 0.7. \end{aligned} \tag{6.2}$$

It is clear that these can be satisfied by a number of specifications lying between two extremes:

$$p(e_{11}) = 0.5, p(e_{10}) = 0, p(e_{01}) = 0.2, p(e_{00}) = 0.3, \tag{6.3}$$

$$p(e_{11}) = 0.2, p(e_{10}) = 0.3, p(e_{01}) = 0.5, p(e_{00}) = 0. \tag{6.4}$$

If (6.3) is the case, then we cannot improve on the overall success rate 0.7 envisaged by the recommendation, whereas, if (6.4) is the case, and if we could identify patients in the various groups, we could clearly attain complete success with all patients. We ought therefore to investigate the patients in the four treatment x response categories to discover whether there are any features that distinguish among them. For example if the distributions of the sexes M and F were

	t_1	t_2
Success	50F	50M
		20F
Failure	50M	30F

then it would surely be sensible to consider allocating males to treatment t_2 and females to treatment t_1.

Thus we consider shifting the emphasis in clinical trials from the customary question posed 'Which treatment is best?' to 'Which treatment is best for which patient?' In our illustrative example, we see that, for the latter question to be answerable, we require to know for each patient in the clinical trial the triplet

(y, t, w), where y $(\in Y)$ denotes an observation on a set of potentially useful indicating features (in our example, sex), where t $(\in T)$ denotes the treatment assigned in the trial (here, t_1 or t_2) and w $(\in W)$ is an observation on a recognised set R of response features (here, success or failure).

A clinical trial will have achieved its purpose if it provides us with a clear picture of the variability of w for given t and y. For this is simply the quantification of the medical concept of prognosis for a patient in current state y if put on treatment t, a concept that is clearly necessary however implicitly it may remain in the formal decision process. To investigate its possible quantification, we may consider some suitable parametric form, say

$$p(w|t, y, \psi) \quad (\psi \in \Psi), \tag{6.5}$$

for the prognosis distributions, and use the data

$$v = \{(y_i, t_i, w_i): i = 1, \ldots, n\} \tag{6.6}$$

from the clinical trial to obtain, in the same kind of way as for the diagnostic assessment, the predictive forms of the prognosis distributions

$$p(w|t, y, v). \tag{6.7}$$

There have been some recent attempts to tackle this kind of problem quantitatively in an estimative rather than predictive way (see for example, Peel *et al.*,[16] Hughes *et al.*,[17] Ginsberg and Offensend,[18] Norris *et al.*[19] with w a simple binomial response, survival or death, and t for a single treatment, so that the statistical technique is indistinguishable from estimative diagnosis.

An example of more sophisticated model-building is implicit in the discussion of prognosis and the effects of treatment in Pickering.[20] There the many straight line fits to the plots of cumulative survival percentages against logarithm of time suggest that the basic family of prognostic distributions (6.5), with w logarithm of survival time and y a measure of current blood pressure, are well characterised by normal distributions with mean $\alpha_t + \beta_t y$ and variance σ_t^2. It is easy to visualise circumstances circumstances in which it will be desirable to give different treatments at different current blood pressure levels; for example, if the σ_t are the same but the β_t different.

7. *MISCELLANEOUS PROBLEMS OF A SYNTHESIS*

The main obstacle to any attempt to bring together the various strands of clinical decision making is the difficulty of obtaining any clearcut picture of the clinician's (or the patient's and clinician's) assessment of the advantages and disadvantages of moving from a current state, assumed to be described by y, to a future state, assumed to be measured by w. Allowance must also be made for the cost of treatment, not only in monetary terms of treatment materials and manpower but as importantly in terms of discomfort to the patient. It is natural that clinicians must show reticence over such matters, for some decisions inevitably call for some assessment on the value of life itself, especially in problems where there is competition between patients for limited equipment, currently, for example, in renal dialysis and transplants. Yet decisions are undoubtedly made and imple-

mented, however implicitly any preference or utility structure enters into the decision making process. If a utility structure, of the type $U(\kappa, y, t, w)$, envisaged in (2.7), is assumed then one sensible way to arrive at a decision is through the expected utility analysis of (2.8). In some areas of medicine, attempts are being made to place even monetary assessments on certain states, for example on the extent of industrial disease, such as degree of pneumoconiosis, or of injury such as thalidomide deformity. We hope that we may make some direct headway in a cooperative effort to arrive at some utility structure for Conn's syndrome (section 4).

In the absence of any examples of direct assessment, we can turn our attention only to ways in which the decision theorist may make some contribution to improving the effectiveness of treatment. One such way is to ask whether it is possible to uncover the implicit utility structure of a clinician by observing his actions or decisions in practice. Aitchison[21] has discussed the feasibility of such reconstruction with different qualities of data, and here we comment on a simple application, reported by Aitchison *et al.*[22]

The treatment of thyrotoxicosis can take one of three forms, anti-thyroid drug therapy, surgery or radio-iodine therapy, and at present there appears to be no method of determining with any certainty the appropriate treatment for a particular case. The question is therefore whether we can discern any inconsistencies in the allocations of treatment made by clinicians. In the study referred to, the clinicians agreed that their allocations were affected by only five features—age, goitre size, an agreed index of general health, and two technical tests of thyroid function. Each clinician was presented with these feature vectors for 40 cases and asked to state what treatment he would advise. If we regard the three treatments as the three categories of the set C we see that a clinician's allocations effectively define a set of diagnostic case records. The extent to which it is possible to simulate the clinician is a measure of his consistency. Aitchison *et al.*[22] show that a suitable measure of inconsistency is the average degree of uncertainty associated with the predictive assessment. In this assessment, any case in which the predictive assessment differs from the actual allocation by the clinician is treated as if it were at the point of maximum uncertainty.

It is shown that the clinicians differ considerably in their inconsistency measures. It is further shown, by a technique similar to the feature selection aspect of section 5, that effective use is made of only the first three components of y and the method of use varies from clinician to clinician. It is hoped that revelation of the detailed nature of these differences can help clinicians to formulate the decision problem in more formal terms and to arrive at some greater consensus. After feedback of the analysis to the clinicians, a further study of treatment allocations is in progress.

We touch on a few further aspects of the synthesis of decision making in clinical medicine by posing questions to which we provide only partial answers. Should treatment be allocated on the basis of the most plausible disease category? The answer is clearly 'Not necessarily'. It is a well known feature of statistical decision theory that it is not necessarily optimum to take that action which is associated with the most plausible state of nature. In terms of (2.8), the maximising t for the double sum will not necessarily be the same as that maximising the single sum associated with the most plausible category, k say. In view of this answer, we may then be inclined to ask a more far-reaching question.

To what extent is diagnosis necessary? The patient is largely unconcerned with

the technical name given to his particular condition; his preoccupation is undoubtedly to move from an unpleasant current state y to a happier state w with as little discomfort to his body or his purse as possible. For the patient, therefore, the utility in (2.8) takes the form $U(y, t, w)$, independent of κ. In evaluating (2.8), therefore, we could then take $U(y, t, w)$ to the left of the inner accumulation and so rewrite (2.8) as

$$\sum_{w \in W} u(y, t, w)\, p(w|t, y, v), \tag{7.1}$$

where effectively $p(w|t, y, v)$ is the predictive form of the prognosis distribution discussed at (6.7). Thus conceptually the diagnostic phase of the decision process is unnecessary. There is indeed some evidence that this is an approach which is becoming more and more acceptable in medicine; see for example, Knill-Jones *et al.*[23]

How does the Hippocratic oath affect clinical decision making? The first aspect of this question is really a continuation of a previous one. Many approaches to the diagnostic phase of clinical medicine have been obsessed by the objective of minimising the misdiagnosis rate. This criterion corresponds to a utility structure which sets equal the losses associated with all the different possible kinds of misdiagnosis. This is clearly a very special kind of utility structure, and to the extent that the true but implicit structure differs from it this statistical approach is inadequate. Until such time as a clinician is prepared to state explicitly the utility structure involved, the role of the decision theorist must be confined to advising on diagnostic assessment, prognosis assessment, and uncovering implicit utility structures.

The second aspect concerns the tension between the roles of doctor and scientist within the clinician. Denote by $G(\kappa)$, $G(\kappa, \alpha)$, ... the expected gain of information concerning κ, (κ, α), ... from the observation of a set of features on a new patient. Then on the basis of the assumptions h of section 3, and in particular using (3.1), we can show that

$$G(\kappa, \alpha) = G(\kappa) \tag{7.2}$$

$$= G(\alpha) + \int_A G(\kappa|\alpha)\, p(\alpha) d\alpha, \tag{7.3}$$

$$G(\kappa, \alpha, \theta) = G(\kappa) + \sum_{\kappa \in C} G(\theta|\kappa)\, p(\kappa). \tag{7.4}$$

Now in the diagnostic phase, the only unknown parameter relevant to treatment of the new patient is κ, and so, in choosing to observe features on him, we are concerned with maximisation of $G(\kappa)$. The relation (7.2) shows that the choice which seeks to obtain maximum information concerning both κ *and* α is equivalent to one that takes the Hippocratic action of maximising information concerning κ alone; whereas (7.3) and (7.4) demonstrate that maximisation of expected information concerning α alone, or of expected information concerning θ, from a patient without regard to κ are unhippocratic.

There are many obvious defects and oversimplifications in the approach to medical decision making that we have presented here. Because of the difficulty of persuading clinicians to declare themselves sufficiently to allow some form of utility structure we have emphasised those aspects of their decision making that are most accessible, mainly the diagnostic aspect. By setting out principles of diagnosis carefully, we have demonstrated that, in practice, large differences do arise in the

employment of different methods, and that performance in diagnosis differs widely in different individuals and falls short of reasonably objective criteria. Clearly the observation of features is an attempt to gain information about the category of an individual problem. However, the observation of features has associated costs, not only in financial terms but also in terms of patient discomfort and delay before treatment. In considering the diagnostic phase in isolation, how do we find common units for the measurement of information gain and such costs? Moreover, when we realise that the decision making process is taking place in time and calls for sequential actions, we are faced with the awkward question of when to stop the diagnostic phase and proceed seriously to treatment? Moreover, in the process of treatment, there are in practice opportunities of adjustment, even of returning to the diagnostic phase because of information obtained in the process of treatment. In short, to describe the decision making process completely, we require something more akin to a control process. We are clearly a long way from achieving this even in simple cases.

8. REFERENCES

1. Ledley, R. S. and Lusted, L. B., Medical Diagnosis and Modern Decision Making, in *Mathematical Problems in the Biological Sciences, Proceedings of Symposia in Applied Mathematics,* 1962, 14, pp. 117-158.
2. Lusted, L. B., *Introduction to Medical Decision Making,* Thomas, Springfield, Illinois, 1968.
3. Aitchison, J. and Kay, J. W., Estimative and Predictive Diagnosis: A Critical Comparison. Submitted to *Biometrika,* 1974.
4. Geisser, S., Posterior Odds for Multivariate Normal Classifications, *J.R. Statist. Soc. B,* 1964, 26, 1, pp. 69-76.
5. Geisser, S., The Inferential Use of Predictive Distributions, in Godambe, V. P., and Sprott, D. A., (eds.), *Foundations of Statistical Inference,* Holt, Rinehart and Winston, Toronto, 1972.
6. Guttman, I. and Tiao, G. C., A Bayesian Approach to some Best Population Problems, *Ann. Math. Statist.,* 1964, 35, pp. 825-835.
7. Aitchison J. and Sculthorpe, D., Some Problems of Statistical Prediction, *Biometrika,* 1965, 52, pp. 469-483.
8. Zellner, A. and Chetty, V. K., Prediction and Decision Problems in Regression Models from the Bayesian Point of View, *J. Amer. Statist. Ass.,* 1965, 60, pp. 608-616.
9. Dunsmore, I. R., A Bayesian Approach to Classification, *J.R. Statist. Soc. B,* 1966, 28, 3, pp. 568-577.
10. Dunsmore, I. R., A Bayesian Approach to Calibration, *J.R. Statist. Soc. B,* 1968, 30, 2, pp. 396-405.
11. Dunsmore, I. R., Regulation and Optimisation, *J.R. Statist. Soc. B,* 1969, 31, 1, pp. 160-170.
12. Lindley, D. V., The Choice of Variables in Multiple Regression, *J.R. Statist. Soc. B,* 1968, 30, 1, pp. 31-66.
13. Dixon, W. J., *BMD Biomedical Computer Programs,* University of California Press, 1970.
14. Lindley, D. V., On a Measure of the Information Provided by an Experiment, *Ann. Math. Statist.,* 1956, 27, pp. 986-1005.
15. Taylor, T. R. *et al.*, Doctors as Decision Makers: a Computer-Assisted Study

of Diagnosis as a Cognitive Skill, Brit. Med. J., 1970, 3, pp. 35-40.
16. Peel, A. A. F. *et al., Brit. Heart J.*, 1962, 24, pp. 745-760.
17. Hughes, W. L. *et al.*, Myocardial Infarction Prognosis by Discriminant Analysis, *Archives of Internal Medicine*, 1963, 111, pp. 338-345.
18. Ginsberg, A. S. and Offensend, F. L., An Application of Decision Theory to a Medical Diagnosis-Treatment Problem, *I.E.E.E. Transactions of Systems Science and Cybernetics*, 1968, SSC4, pp. 355-362.
19. Norris, R. M. *et al.*, A New Coronary Prognostic Index, Lancet, 1969, 1, pp. 274-281.
20. Pickering, G., *High Blood Pressure*, Churchill, London, 1968.
21. Aitchison, J., Statistical Problems of Treatment Allocation. *J.R. Statist. Soc. A*, 1970, 133, 2, pp. 206-238.
22. Aitchison, J. *et al.*, Consistency of Treatment Allocation in Thyrotoxicosis, *Quart. J. Medicine*, 1973.
23. Knill-Jones, R. P. *et al.*, The Use of a Sequential Bayesian Model in the Diagnosis of Jaundice by Computer, to be published, 1973.

12 PARAMETRIC MODELLING IN STATISTICAL DIAGNOSIS

12.1 Introduction

Much of the attention of statisticians in applications in medicine has been directed towards problems in medical research such as clinical trials. There are equally important problems arising from the uncertainties and variabilities in the clinical management of individual patients. Aitchison and Kay (15:1975) examine this process of patient management recognising and modelling various aspects of the problem such as past experience of the clinician, observation and measurement, diagnosis, prognosis, treatment allocation and assessment. There is little point in going into detail in this brief commentary. Greater quantification of the practice of clinical medicine must come, and will require the careful collection of data to allow the fitting of the various aspects suggested in this paper. In this section the discussion is confined to the main aspects of parametric modelling of more complex problems arising directly out of consultation in statistical diagnosis.

12.2 Statistical diagnosis when basic cases are not classified with certainty

The purpose of Aitchison and Begg (16:1976) is to provide a generalisation of statistical diagnostic techniques to situations where the basic data set consists of cases for each of which a feature vector x is available together with a vector u of the relative plausibilities of the possible types. The need for such a generalisation had been seen in a reexamination of the problem

of the differential diagnosis of Conn's syndrome (Ferriss et al, 1970), for which the training set was reduced from 34 to 31 cases because of some uncertainty, even at histopathology, of the types of three cases.

In order to model this situation a number of new concepts and techniques had to be introduced.

1. In straightforward diagnostic systems the diagnostic statement is *simple* in the form 'this case is of type t'. To model the pattern of variability of such statements we require distributions over the set T of possible disease types. When a basic case is classified with uncertainty the diagnostic statement is *composite* itself, being a probability distribution $\{u_t : t \in T\}$ over T. To model the pattern of variability of such statements we thus require distributions over the class of distributions over T, one step further up the hierarchy of distributional statements: simple, composite, distributional.

2. The modelling of the variability of (u,x) thus requires parametric modelling of distributions on a space of the form $S^c \times R^d$, where S^c is the positive simplex

$$\{u : u_i > 0 \quad (i = 1,\ldots,c) : u_1 + \ldots + u_c < 1\}$$

and R^d is d-dimensional real space. This is an awkward intractable space, which, however, can be converted to an analysable problem in R^{c+d} by use of the one-to-one mapping between S^c and R^c through the loglinear, logistic transformations

$$v_t = \log\{u_t/(1 - u_1 - \ldots - u_c)\} \quad (t = 1,\ldots,c),$$

$$u_t = \exp(v_t)/\{(1 + \exp(v_1) + \ldots + \exp(v_c)\} \quad (t = 1,\ldots,c).$$

We shall discuss further justification and developments of this transformation in §13.

3. The transformed data $D = \{(v_i, x_i) : i = 1, \ldots, n\}$ are then analysed through the tools of multivariate regression analysis in its predictive form. Some of the predictive distribution results here involve conditional multivariate Student distributions and are extension of the results of Aitchison and Dunsmore (13:1975).

4. For a new case with feature vector y the analysis leads directly to an assessment in the form $p(v|y,D)$, where v is related to the composite diagnosis u through the logistic transformation. There remains therefore the problem of reducing this essentially distributional diagnostic statement to a more usable and interpretable composite statement. Conditional probability arguments then lead to the evaluation of certain integrals. For example, for the case of two types 1 and 2 we have

$$p(t = 1|y,D) = \int_{-\infty}^{\infty} \frac{e^v}{1+e^v} p(v|y,D)\,dv.$$

Approximate methods for the evaluation of such integrals are provided. Lauder (1978) discusses further the computational problems associated with integrals of this type.

Aitchison and Begg (16:1976) had no data to illustrate their method, which was developed to show that composite diagnostic statements can be accepted as a basis for the construction of diagnostic systems, with a view to encouraging clinicians to use composite statements if they are at all uncertain. The method has since been applied to problems of differential diagnosis of

auditory malfunction and of pre-eclampsia by Begg (1976).

12.3 The system transfer problem in statistical diagnosis

The basic problem considered by Aitchison (18:1977) is one of modelling, how to link a diagnostic system devised from data

$$D_1 = \{(t_i, x_i) : i = 1, \ldots, n\}$$

from clinic 1, through independent calibrative data

$$C_{12} = \{(y_{1j}, y_{2j}) : j = 1, \ldots, k\}$$

between clinics 1 and 2, to produce a diagnostic system in clinic 2. Aspects of importance are the following.

1. The assumptions of the parametric modelling are set out carefully to take account of the nature of the data. This leads to the eventual use of the predictive density function

$$p(x_2|t, C_{12}, D_1) \propto \int_{X_1} p(x_2|x_1, C_{12}) p(x_1|t, D_1),$$

where $p(x_2|x_1, C_{12})$ is the calibrative distribution from clinic 1 measurement to clinic 2 measurement and $p(x_1|t, D_1)$ is the predictive distribution on which clinic 1 diagnosis is based.

2. The discrepancy between the use of 'naive calibration', which uses a calibrated point estimate obtained by inverse regression, and the calibrative diagnostic assessment described in (1), is analysed for the case of perfect information where all the distributions are known with precision, in an attempt to locate the sources of these discrepancies. In particular, factors associated with the discriminating power of the diagnostic features, the unreliability of the calibration process and the standardised

distance of the naive calibrate from the centre of the feature distributions, are identified. The case of perfect information, where the naive calibration method is certainly wrong, allows us to explore the possibility of substantial differences between the full and the naive calibrative diagnostic methods. Two illustrative examples, quite possible in practice, show naive odds of 28 to 1 and 20,000 to 1 reduced to 9 to 1 and 25 to 1 respectively.

3. The detailed distribution theory required for the normal case is developed.
4. Modifications are made for the case of partial calibration, where only some of the diagnostic features require calibration.
5. The technique is applied to the differential diagnosis of Conn's syndrome for new cases, for which the method of measurement of one of the features has changed from that of the training set. Out of 43 new cases eight are found to have diagnostic assessments which differ so much between the predictive calibrative and the naive calibrative methods that they lead to substantial practical differences in patient management.

The method is readily applied and the potential misrepresentations in ignoring the effect are so substantial that there is no real excuse for not taking full account of the need to calibrate when the occasion demands.

12.4 The clinic amalgamation problem in statistical diagnosis

Aitchison (19:1979) develops methods for a more complex calibrative-diagnostic problem in which two or more clinics wish to pool their diagnostic data in order to construct a more reliable

diagnostic system than any one clinic could produce by itself.

The objective here is to provide each clinic with a diagnostic system with its own method of measurement. When only two clinics are involved the data consist of two diagnostic training sets

$$D_i = \{(t_{ij}, x_{ij}) : i = 1, \ldots, n_i\} \quad (i = 1, 2)$$

and a calibrative set

$$C_{ij} = \{(y_{1j}, y_{2j}) : j = 1, \ldots, n\}.$$

There are many possible ways of modelling and which is chosen must depend on the nature of the data. Only the main features of one form of modelling need be reported to bring out new developments of the approach. The case of two types only is used, and the presentation is for the case where the calibration experiment allows calibration from clinic 2 to clinic 1.

1. The 'diagnostic paradigm' is adopted, modelling the conditional distribution of type t for given feature vector x by

$$p(t = 1 | x, \delta) = \Phi(\delta^T x).$$

A new feature here is the use of the normal rather than the popular logistic distribution function. The reason for this is that it yields explicit forms, through convolution integrals, for the diagnostic model for clinic 2.

2. If the calibration model is set up in terms of a conditional multivariate normal density function

$$p(x_1 | x_2, \gamma) = \phi(x_1 | Ax_2, B)$$

then the diagnostic model for clinic 2 has a parametric form indexed by $\gamma = (A, B)$ and δ:

$$p(t=1|x_2,\gamma,\delta) = \int_{X_1} \Phi(\delta^T x_1)\phi(x_1|Ax_2,B)\,dx_1$$
$$= \Phi(\varepsilon^T x_2),$$

where $\varepsilon = A^T\delta/\sqrt{(1+\delta^T B\delta)}$, so that the induced diagnostic paradigm for clinic 2 is also of the normal linear form.

3. This completion of the model building allows us to write down the likelihood and then to obtain an approximation to the posterior distribution $p(\gamma,\delta|C_{12},D_1,D_2)$ through Bayesian maximum likelihood theory, with a view to adopting a predictive approach towards the assessment of new cases.

4. To illustrate this assessment suppose that a new patient with feature vector x_1 is to be diagnosed in clinic 1. Suppose that the marginal in δ of the posterior distribution is $\phi(\delta|d,G)$. Then

$$p(t=1|x_1,C_{12},D_1,D_2) = \int_\Delta \Phi(\delta^T x_1)\phi(\delta|d,G)\,d\delta$$
$$= \Phi\left\{\frac{\delta^T x_1}{\sqrt{(1+x_1^T G x_1)}}\right\}.$$

When the calibration experiment is a natural one diagnostic assessments for new cases in clinic 2 take a similar form. When the calibration experiment allows only calibration from x_2 to x_1 then for new cases in clinic 2 the diagnostic assessment for the normal linear model takes a more complicated form

$$pr(t=1|x_2,C_{12},D_1,D_2) = \int_\Gamma\int_\Delta \Phi\left\{\frac{\delta^T A^T x_2}{\sqrt{(1+\delta^T B\delta)}}\right\}\phi(\gamma,\delta|c,d;J)\,d\gamma\,d\delta$$

where $\gamma = (A,B)$. This multiple integral requires for its evaluation numerical or Monte Carlo methods, reinforcing the

soundness of the advice to perform a natural calibration experiment wherever possible.

5. For a simple illustrative example this diagnostic system is compared with three alternative methods, termed the single relevant clinic method, the system transfer method and the naive calibration method. All of these methods are open to the criticism of neglecting some aspect of the information available and it is again shown that substantial departures occur between the clinic amalgamation system and these other methods.

6. The methods are applied to the amalgamation of two clinics for the differential diagnosis of Conn's syndrome.

12.5 Statistical diagnosis from imprecise data

In the actual construction of a statistical diagnostic system imprecision in the feature vectors, although often recognised, is seldom taken into account. The reasons for the neglect of imprecision in general are presumably

(i) the lack of appropriate statistical methods in this kind of discriminant analysis, and

(ii) the assumption that disregard of such imprecision has negligible consequences in practice.

The purpose of Aitchison and Lauder (20:1979) is to remedy (i) and to investigate (ii).

The main features of the modelling can be briefly described as follows.

1. Model diagnosis is expressed in diagnostic paradigm form with the conditional distribution of type t on accurate feature vector x given in parametric form $p(t|x,\delta)$.

2. Suppose that we cannot observe x but only a possibly inaccurate vector y. Suppose, however, that we know from experience the nature of this imprecise measurement process and can in fact specify the conditional density function $p(x|y)$ of x for given y.

3. The diagnostic model in terms of the observable y is then

$$p(t|y,\delta) = \int_X p(t|x,\delta)p(x|y)dx.$$

4. The likelihood based on a training set

$$D = \{(t_i,x_i) : i = 1,\ldots,n\}$$

can then be obtained, followed by some reasonable approximation to $p(\delta|D)$, so that the diagnostic assessment for a new case with observed feature vector y is given by

$$p(t|y,D) = \int_\Delta p(t|y,\delta)p(\delta|D)d\delta.$$

5. Comparison is then made between adopting logistic and cumulative normal models for $p(t|x,\delta)$, with the analytical advantage going to the normal model because a closed form can then be obtained for the diagnostic assessment integral.

6. Means are provided for the operation of the system, and illustrative applications to a subproblem of the differential diagnosis of Cushing's syndrome and to a solution of the clinic amalgamation problem of Aitchison (19:1979). In particular, for the Cushing problem it is found that as the imprecision coefficient of variation is increased from 0 to its realistic value the application of the Newton-Raphson method becomes more and more difficult until the whole basis

of the maximum-likelihood distributional aspect of $p(\delta|D)$ becomes suspect.

7. The illustrative examples demonstrate that ignoring imprecision can given diagnostic assessments with a false appearance of firmness. Explicit recognition of imprecision can be incorporated in diagnostic modelling and has the expected effect of reducing the firmness of the diagnostic assessments. This effect can be so extreme that diagnostic assessments are practically the same whatever the feature vector. If the imprecision is appreciable standard procedures, such as Newton-Raphson iteration, which work readily under the assumption of precise data, fail completely for the degree of imprecision actually present.

Aitchison (21:1979) examines the theory and practical applications for the structurally similar problem of calibration and assay when the standards have themselves an element of imprecision.

AITCHISON, J. and BEGG, C.B. (1976)

Statistical diagnosis when basic cases are not classified with certainty

Reprinted from *Biometrika* 63, 1-12

Biometrika (1976), 63, 1, *pp.* 1–12
Printed in Great Britain

Statistical diagnosis when basic cases are not classified with certainty

BY J. AITCHISON AND C. B. BEGG
Department of Statistics, University of Glasgow

SUMMARY

A need is identified for statistical diagnostic techniques based on data sets containing cases which have not been allocated to a single diagnostic type with certainty but for which only an assessment of the probabilities of the types is available. In order to construct a statistical diagnostic system applicable to new cases it is necessary to introduce even more complex diagnostic assessments and a central part of the analysis is concerned with the reduction of such assessments to simpler, more interpretable forms. The theory provides a generalization of current statistical diagnostic techniques, and its relationship to these is discussed.

Some key words: Discrimination; Hierarchy of diagnostic statements; Logistic transform; Medical diagnosis; Predictive distribution.

1. INTRODUCTION

Diagnosis is often considered as the allocation of a case to a single type of a finite set $T = \{1, \ldots, r\}$ of possible types on the basis of some observed feature vector x associated with the case. More generally and often more realistically it is the assessment of the relative plausibilities $u_1, \ldots, u_r$ of each of the possible types for the case. Many statistical techniques such as discriminant analysis (Fisher, 1936; Smith, 1947), predictive discriminant analysis (Geisser, 1964; Dunsmore, 1966; Aitchison & Kay, 1975) and logistic discrimination (Cox, 1966; Day & Kerridge, 1967; Anderson, 1972) have been evolved. The insistence of all these techniques that the cases of the data base have a clear typing is a restriction commonly unfulfilled or only artificially satisfied in practice. Such difficulties of firm typing are well recognized in certain areas of medicine, for example in psychiatry, and in other areas they are often side-stepped. For example, in the differential diagnosis of Conn's syndrome (Ferriss *et al.*, 1970), for which there are two types, the basic set of past records was reduced from 34 to 31 because of some uncertainties, even at histopathology, of the types of three of the cases. If such doubtful cases are possible in the future then any statistical diagnostic technique based on the reduced set of 31 past cases could appear deceptively better in theory than its subsequent realization in practice. Attempts to justify the omission of such cases are based on the assumption that the typing has been determined by factors, such as histopathology of adrenal sections in the case of Conn's syndrome, other than the feature vectors, such as plasma concentrations of electrolytes, now under consideration; and that cases doubtfully typed on the basis of the factors will be more clearly separated by the feature vectors. Such reasoning can, however, be circular since the object of using the feature vector is often simply to avoid the use of the factors, for example, in Conn's syndrome, to avoid an unnecessary operation, but to arrive at results which would be attained by their use. Writers often report

higher misclassification rates in application than those predicted from the basic set, even when due allowance is made for resimulation bias (Lachenbruch & Mickey, 1968) and sampling variability (Aitchison & Kay, 1975). The omission of such doubtful cases may well be a cause of this persisting phenomenon.

On the other hand, particularly when past cases are in short supply, there is a temptation to impute firm typing to cases where some diagnostic doubt remains, in order that they meet the requirements of current statistical diagnostic techniques. Any such case, if wrongly typed, is likely to distort diagnosis much more seriously than it would within a system which allows some expression of doubt about its type.

The purpose of this paper is to provide a generalization of statistical diagnostic techniques to situations where the basic data set consists of cases for each of which a feature vector x is available together with a vector u of the relative plausibilities of the possible types. If such techniques were known to be available then diagnosticians in their consideration of the basic cases could be encouraged to express their findings in more realistic terms by stating explicitly their uncertainties.

2. DIAGNOSTIC STATEMENTS

A diagnosis of a case consists of a statement relating the case to one or more of the types of T. Diagnostic statements can take many different forms at different levels of probabilistic sophistication, and recognition of this hierarchy plays an important role in the development of statistical techniques of diagnosis. We therefore first define these forms of diagnoses in ascending order of sophistication.

A simple diagnosis t states that a case is of a specified type $t \in T$.

A composite diagnosis $u = \{u_t : t \in T\}$ states that a case is of type t with probability u_t $(t \in T)$, and is thus simply a probability distribution on the finite set T. We shall restrict u to the strictly positive r-dimensional simplex $S^r = \{u : u_t > 0 \ (t \in T), \Sigma_T u_t = 1\}$ so that a composite diagnosis never completely excludes any of the possible types.

A distributional diagnosis w attaches a plausibility to each possible composite diagnosis $u \in S^r$ and so w is a density function $w = p(u)$ on S^r. Thus w belongs to the class $\mathscr{P}(S^r)$ of density functions on S^r.

We could even envisage a hyperdistributional diagnosis which assigns a probability distribution on $\mathscr{P}(S^r)$, assigning plausibilities to the possible distributional diagnoses. For example, if, for the basic set of cases, each of a panel of diagnosticians had made individual composite diagnoses we might express the panel's view as a distributional diagnosis; the natural statistical expression for the diagnosis of a new case would then be a hyperdistributional diagnosis.

The probabilistic forms of simple and composite diagnoses are obvious. For the representation of distributional diagnoses an immediate question is which parametric forms of density function commend themselves as practically relevant and tractably acceptable.

For many problems the natural class of distributions to consider for a vector confined to the simplex S^r is the Dirichlet class. Such a choice is often dictated by the fact that the vector is a parametric vector of multinomial probabilities and by the convenient mathematical property of conjugacy which the Dirichlet distribution bears to the multinomial distribution; and the choice is usually justified by an appeal to the richness of the Dirichlet class in providing a variety of density functions. The present role of u in S^r is different since u is not an

unknown parameter but a directly observable composite diagnosis. In such circumstances the use of the Dirichlet class is unattractive; no conjugate prior class of density functions exists for its parameters. The basic difficulty with distributional diagnoses is the mathematical awkwardness of handling the restriction of each u_t to the interval (0, 1) and also the summation requirement $\Sigma u_t = 1$. This difficulty can be removed, as in many other statistical contexts, by the convenient one-to-one correspondence between the r-dimensional simplex S^r and $(r-1)$-dimensional Euclidean space R^{r-1}, through the equivalent logistic relationships, for $t = 1, \ldots, r-1$,

$$v_t = \log(u_t/u_r), \tag{2·1}$$

$$u_t = e^{v_t} \Big/ \left(1 + \sum_{i=1}^{r-1} e^{v_i}\right), \quad u_r = 1 \Big/ \left(1 + \sum_{i=1}^{r-1} e^{v_i}\right). \tag{2·2}$$

The mathematical reason for our restriction of u to the positive simplex is now clear since (2·1) is undefined if u has any zero components. Any composite diagnosis $u \in S^r$ can be characterized through (2·1) in terms of its associated vector $v \in R^{r-1}$, and we shall use the term composite diagnosis to refer to both forms u and v; the notation and context will always make clear which form is intended. Similarly a distributional diagnosis can now be a density function for u on S^r or for v on R^{r-1}.

We can then obtain a rich class of density functions to represent distributional diagnoses by taking v-forms which are $(r-1)$-dimensional multinormal distributions. Indeed the associated distributions of u on S^r are richer than the Dirichlet class. Consider first $r = 2$, when such a specification would assign $v = \log\{u/(1-u)\}$ a $N(\mu, \sigma^2)$ distribution. Compare this class with the class of beta density functions $u^{\alpha-1}(1-u)^{\beta-1}/B(\alpha, \beta)$ $(0 < u < 1)$, the Dirichlet class for $r = 2$. The number of parameters is the same, namely 2, and it is easy to verify that the good variety of U-shaped, J-shaped, flattish to sharpish unimodal curves for u furnished by the beta class can be obtained with suitable choice of μ and σ^2 for the normal v-specification. For instance, choice of μ near zero and σ large will yield large positive or large negative v with a resultant U-shaped distribution for u in the interval (0, 1); whereas choice of $\mu = -\log 3$ and small σ will yield a distribution for u sharply peaked at $u = \frac{1}{4}$. Here, for $r > 2$, the number of parameters in the multivariate normal specification is $(\frac{1}{2}r-1)(r+2)$ compared with r in the Dirichlet specification. From this and by arguments similar to those for $r = 2$ the greater richness of the multinormal representation is easily verified.

We emphasize that although no component of u can be unity or zero so that the possibility of diagnosis with certainty is excluded, a simple diagnosis can be effectively achieved by assigning one of the components close enough to unity. In practical terms this means that the basic set could virtually contain cases diagnosed with certainty in addition to composite data.

3. Notation and distributional results

In what follows we shall require a convenient notation for some distributions associated with d-dimensional multinormal models and also some results concerning properties of these distributions, We write:

R^d, d-dimensional real space;

$\mathcal{S}^d$, the space of positive-definite symmetric matrices of order d;

$\Gamma_d(g) = \pi^{\frac{1}{4}d(d-1)}\Gamma(g)\,\Gamma(g-\frac{1}{2})\ldots\Gamma\{g-\frac{1}{2}(d-1)\}$, the Siegel (1935) generalization of the gamma function.

The definitions of the multinormal, Wishart, normal-Wishart and generalized Student distributions are provided in Table 1. The multinormal distribution is specified for convenience in terms of its precision matrix τ, the inverse of the covariance matrix. The dimensions of vectors and matrices are obvious from the context.

Table 1. *Definitions of standard distributions*

Distribution and notation	Sample space	Density function
Normal $\mathrm{No}_d(\mu, \tau)$	$x \in R^d$	$(2\pi)^{-\frac{1}{2}d}\lvert\tau\rvert^{\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)'\,\tau(x-\mu)\}$
Wishart $\mathrm{Wi}_d(\nu, \tau)$	$y \in \mathscr{S}^d$	$\dfrac{\lvert\frac{1}{2}\tau\rvert^{\frac{1}{2}\nu}\lvert y\rvert^{\frac{1}{2}(\nu-d-1)} \exp\{-\frac{1}{2}\mathrm{tr}\,(\tau y)\}}{\Gamma_d(\frac{1}{2}\nu)}$
Normal-Wishart $\mathrm{NoWi}_d(b, c, g, h)$	$(x, y) \in R^d \times \mathscr{S}^d$	$p(x\lvert y)$ is $\mathrm{No}_d(b, cy)$ $p(y)$ is $\mathrm{Wi}_d(g, h)$
Student $\mathrm{St}_d(k, b, c)$	$x \in R^d$	$\dfrac{\Gamma\{\frac{1}{2}(k+1)\}}{\pi^{\frac{1}{2}d}\Gamma\{\frac{1}{2}(k-d+1)\}\,\lvert kc\rvert^{\frac{1}{2}}\{1+(x-b)'\,(kc)^{-1}\,(x-b)\}^{\frac{1}{2}(k+1)}}$

We shall also require marginal and conditional distributions associated with the generalized Student distribution. Suppose that x is $\mathrm{St}_d(k, b, c)$ and that we partition x into (x_1, x_2), where x_1 and x_2 have dimensions d_1 and $d_2 = d - d_1$. Let the corresponding partitioning of b, c and $\tau = c^{-1}$ be

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}, \quad \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}.$$

Then $p(x_1)$ is $\mathrm{St}_{d_1}\{k-d_2, b_1, k(k-d_2)^{-1}c_{11}\}$ or $\mathrm{St}_{d_1}\{k-d_2, b_1, k(k-d_2)^{-1}(\tau_{11}-\tau_{12}\tau_{22}^{-1}\tau_{21})^{-1}\}$; $p(x_2|x_1)$ is

$$\mathrm{St}_{d_2}[k, b_2+c_{21}c_{11}^{-1}(x_1-b_1), (c_{22}-c_{21}c_{11}^{-1}c_{12})\{1+k^{-1}Q(x_1)\}]$$

or

$$\mathrm{St}_{d_2}[k, b_2-\tau_{22}^{-1}\tau_{21}(x_1-b_1), \tau_{22}^{-1}\{1+k^{-1}Q(x_1)\}], \tag{3·1}$$

where

$$Q(x_1) = (x_1-b_1)'\,c_{11}^{-1}(x_1-b_1) = (x_1-b_1)'\,(\tau_{11}-\tau_{12}\tau_{22}^{-1}\tau_{21})\,(x_1-b_1). \tag{3·2}$$

4. A DIAGNOSTIC MODEL FOR MULTINORMAL COMPOSITE DATA

We can use the terminology of diagnostic statements to describe data sets. Thus past case records of the form (t, x) with $t \in T$ constitute simple data, whereas past case records of the form (u, x) with $u \in S^r$, or (v, x) with $v \in R^{r-1}$, constitute composite data. To construct a statistical model for diagnosis based on a composite data set we have to make assumptions about the probabilistic mechanism which generates case records. In § 2 we suggested the multinormal as a suitably flexible class of descriptions of the marginal distribution of v. If the feature vector space is R^f and if the distribution of the feature vector x is multinormal then the natural additional assumption to consider is that the distribution of a case record (v, x) is also multinormal. We shall in this section follow through the diagnostic consequences of this basic assumption. In any potential application we would have to examine the validity

of the assumptions. For example, the feature vector may require transformation to satisfy the multinormal requirement; if some of the features are categorical in form, such as headache/no headache, then the multinormal assumption is clearly not valid.

We now make explicit the assumptions of the model.

Assumption 1. The distribution of any case record (v, x) is $\mathrm{No}_{r-1+f}(\mu, \tau)$.

Assumption 2. The distributions of any finite set of case records $(v_1, x_1), \ldots, (v_n, x_n)$ are independent.

Assumption 3. The prior distribution of (μ, τ) is $\mathrm{NoWi}_{r-1+f}(b, c, g, h)$.

Assumption 4. The data z available consist of case records $z_1 = (v_1, x_1), \ldots, z_n = (v_n, x_n)$.

Assumption 5. The new case under consideration, with unknown composite diagnosis $v \in R^{r-1}$ and known feature vector $x \in R^f$, has arisen from exactly the same probabilistic mechanism as the past case records $(v_1, x_1), \ldots, (v_n, x_n)$.

For a fuller account of similar diagnostic assumptions with simple data and their relevance to clinical medicine, see Aitchison & Kay (1975), Aitchison & Dunsmore (1975, §11·2). The advantages of realism of the resulting predictive diagnosis over the estimative rivals are indicated by Aitchison (1975) and will be more fully presented elsewhere.

Note that Assumption 1 implies that the marginal distribution of feature vectors, that is over all cases, is multivariate normal. For those accustomed to visualizing diagnostic feature vectors as separating into clusters, one for each of the types, this may seem a heretical suggestion. The view of separate clusters is often encouraged by such factors as the exclusion of doubtful cases and by deliberate overrepresentation of rare types to allow satisfactory estimation of parameters. In a reexamination of the data for Conn's syndrome and some other disease groupings we have found no evidence against marginal multinormality. In situations where the feature vector distribution is polymodal the techniques described below are clearly inappropriate; for the necessary adjustment to allow for polymodality see §6.

By adopting Assumption 5 we are envisaging a basic set of composite data produced by a natural process which we anticipate is also going to operate for new cases. This assumption thus excludes from present consideration the situation where the data base has been designed. For example, if some composite diagnoses are rare the clinician may have gone out of his way to collect more cases of these composite diagnoses than would arise naturally in his own clinic. We shall consider what adjustments are necessary for a designed data base in §7.

On the basis of Assumptions 1, 2 and 4 and standard multinormal theory we have that

$$m = \begin{bmatrix} \bar{v} \\ \bar{x} \end{bmatrix}, \quad S = \begin{bmatrix} S_{vv} & S_{vx} \\ S_{xv} & S_{xx} \end{bmatrix},$$

where

$$\bar{v} = \sum_{i=1}^{n} v_i/n, \quad \bar{x} = \sum_{i=1}^{n} x_i/n,$$

$$S_{vv} = \sum_{i=1}^{n} (v_i - \bar{v})(v_i - \bar{v})', \quad S_{vx} = \sum_{i=1}^{n} (v_i - \bar{v})(x_i - \bar{x})', \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})',$$

are sufficient for (μ, τ) and, moreover, independently distributed as $\mathrm{No}_{r-1+f}(\mu, n\tau)$ and $\mathrm{Wi}_{r-1+f}(n-1, \tau)$. It follows (Aitchison & Dunsmore 1975, Table 2·3) from Assumption 3

that the posterior density function $p(\mu,\tau|z)$ of (μ,τ) given the data base z is equal to $\mathrm{NoWi}_{r-1+f}(B,C,G,H)$, where

$$B=C^{-1}(cb+nm),\quad C=c+n,\quad H=h+S+cn(c+n)^{-1}(m-b)(m-b)',$$

$$G=\begin{cases}g+n & (c>0),\\ g+n-1 & (c=0).\end{cases}$$

Hence, on integrating out (μ,τ), we obtain the predictive density function $p(v,x|z)$ for the new case record (v,x) given z as

$$p(v,x|z)=\int_{R^{r-1+f}}\int_{\mathcal{S}^{r-1+f}}p(v,x|\mu,\tau)\,p(\mu,\tau|z)\,d\mu\,d\tau$$

$$=\mathrm{St}_{r-1+f}(G,B,D),$$

where $D=(1+C^{-1})H/G$.

We can immediately arrive at a distributional diagnosis $p(v|x,z)$ by obtaining the appropriate conditional distribution from this joint distribution of (v,x). With an obvious partitioning of B and D,

$$B=\begin{bmatrix}B_v\\ B_x\end{bmatrix},\quad D=\begin{bmatrix}D_{vv} & D_{vx}\\ D_{xv} & D_{xx}\end{bmatrix},$$

we have from (3·1), (3·2) and the above formulae that

$$p(v|x,z)=\mathrm{St}_{r-1}[G,B_v+D_{vx}D_{xx}^{-1}(x-\bar{x}),(D_{vv}-D_{vx}D_{xx}^{-1}D_{xv})\{1+G^{-1}Q(x)\}],\qquad(4\cdot1)$$

where $Q(x)=(x-\bar{x})'D_{xx}^{-1}(x-\bar{x})$. The density function (4·1) provides a measure of the plausibility of the possible composite diagnoses for a new patient with feature vector x and based on the experience embodied in the past records z and the prior information on μ and τ.

In standard statistical diagnostic techniques based on simple data the natural diagnostic statement for a new case is composite in form. Here, starting from composite data our statistical analysis arrives at a distributional diagnosis for a new case. This transition from a data set of one form to a diagnostic statement for a new case of the next higher form in the hierarchy of §2 is a consequence of the need to make probabilistic assumptions about the generation of the data of the basic set. This hierarchial phenomenon leads naturally to the subject matter of §5.

5. Reduction of diagnostic statements

Although a distributional diagnosis is the natural statistical expression of diagnostic assessment based on composite data it is too sophisticated a statistical idea for practical use, except possibly for $r=2$ where the distribution can be presented graphically. The practical way to express a diagnosis under uncertainty is through a composite diagnosis. Since a composite diagnosis is itself a probability distribution on the set T the reduction of a distributional diagnosis, say $p(v|x,z)$, to a composite diagnosis, say $p(t|x,z)$, is a straightforward application of distributional calculus:

$$p(t|x,z)=\int_{R^{r-1}}p(t|v)\,p(v|x,z)\,dv,\qquad(5\cdot1)$$

where for $t = 1, \ldots, r-1$

$$p(t|v) = e^{v_t} \Big/ \left(1 + \sum_{i=1}^{r-1} e^{v_i}\right), \quad p(r|v) = 1 \Big/ \left(1 + \sum_{i=1}^{r-1} e^{v_i}\right),$$

as in (2·2). This natural reduction of a distributional diagnosis to a composite diagnosis thus involves an integration problem. It is easy to see that this natural reduction leads to the mean of the u-distributional diagnosis $p(u|x,z)$ corresponding to the v-distributional diagnosis $p(v|x,z)$.

Our motivating example for the development of the techniques of this paper has been the differential diagnosis of Conn's syndrome for which $r = 2$, and we shall study the natural reduction (5·1) fully for this situation only. The main method we consider for evaluating (5·1) for $r = 2$ extends, however, to provide a practical tool of natural reduction for $r = 3$ and $r = 4$, the other main cases of practical importance, through the use of tables of bivariate and trivariate normal integrals. We shall here only indicate results for these more complex problems, preferring to postpone a full discussion until the completion of the testing of more direct algorithms of numerical evaluation.

For $r = 2$ the integral (5·1),

$$\int_{-\infty}^{\infty} \frac{e^v}{1+e^v} p(v|x,z)\, dv, \tag{5·2}$$

where $p(v|x,z)$ is $\mathrm{St}(k,b,c)$, can be readily evaluated by numerical integration with only modest computing facilities. Our own computations used a Simpson rule method over the range $b - a_k\sqrt{c}$, $b + a_k\sqrt{c}$ where a_k is chosen such that all except ϵ of $\mathrm{St}(k,0,1)$ is contained in $(-a_k, a_k)$, the order of accuracy required. There is however an excellent approximation expressible in terms of the standard normal distribution function Φ, and it is this form of approximation which can be extended to $r = 3, 4$.

The rationale of the approximation is as follows. The function $e^v/(1+e^v)$ is the distribution function of a logistic random variable, say L, and $p(v|x,z)$ is the density function of a $\mathrm{St}(k,b,c)$ random variable, say S. Hence the convolution integral (5·2) can be expressed as the distribution function of $L - S$, evaluated at 0. Now, as pointed out by Cox (1970, pp. 27–8), L can be extremely well approximated by a zero-mean normal variable with a suitably selected quantile in agreement with L. For our purposes agreement of 90 % quantiles seems empirically best, so that L is approximately $N(0, 2{\cdot}942)$. Similarly Student random variables can be well approximated by suitably selected normal random variables and an adequate approximation is obtained by taking S to be approximately $N\{b, kc/(k-2)\}$, that is with mean and variance in agreement. Note that this latter approximation is simply a computational device and not a reversion to the estimative diagnostic method criticized by Aitchison & Kay (1975) which would use a smaller normal variance. Then $L - S$ is approximately $N\{-b, 2{\cdot}942 + kc/(k-2)\}$, so that (5·2) is approximated by

$$\Phi[b/\sqrt{\{2{\cdot}942 + kc/(k-2)\}}]. \tag{5·3}$$

We now indicate the extension of the approximation (5·3) to $r = 3$. We have, for example,

$$p(t=1|x,z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{v_1}}{1+e^{v_1}+e^{v_2}} p(v|x,z)\, dv_1\, dv_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{1+e^{-w_1}+e^{-w_2}} p(w|x,z)\, dw_1\, dw_2$$

on making the substitution $w_1 = v_1$, $w_2 = v_1 - v_2$, written in matrix form as $w = Mv$, say. The first factor of the integrand is now conveniently expressed as the distribution function of a two-dimensional logistic vector, and $p(w|x,z)$ is a $\mathrm{St}_2(k, Mb, McM')$ density function.

Table 2. *Comparison of natural reduction and two approximations for* $k = 20$

			Approximations	
b	c	Natural reduction	(5·3)	(5·5)
8	2	0·999	1·000	0·999
4	2	0·956	0·961	0·963
2	2	0·812	0·811	0·792
1·0	2	0·673	0·670	0·630
0·2	2	0·536	0·535	0·522
8	4	0·997	0·998	0·999
4	4	0·927	0·929	0·944
2	4	0·771	0·769	0·703
1·0	4	0·645	0·644	0·529
0·2	4	0·530	0·529	0·495
8	6	0·993	0·995	0·999
4	6	0·901	0·902	0·925
2	6	0·743	0·741	0·614
1·0	6	0·628	0·627	0·428
0·2	6	0·526	0·526	0·468

The introduction of bivariate normal approximations for these distribution and density functions then leads by arguments similar to those used above to an approximation in terms of a bivariate normal integral over the negative quadrant in the w plane.

An alternative and simpler approximation to natural reduction may be termed modal reduction since it involves approximation to the integral (5·1) by the use of a Taylor expansion of $p(t|v)$ about the mode $\hat{v}$ of the distributional diagnosis $p(v|x,z)$. Since for $p(v|x,z)$ of Student form the mean and mode coincide, the approximation can be shown to be

$$p(t|\hat{v}) + \tfrac{1}{2}\,\mathrm{tr}\{\mathscr{V}(v|x,z)\,D^2 p(t|\hat{v})\}, \tag{5·4}$$

where $\mathscr{V}$ denotes covariance matrix and where D^2 denotes matrix of second-order partial derivatives with respect to the components of v. For $r = 2$, with $p(v|x,z) = \mathrm{St}(k, b, c)$ the approximation (5·4) becomes

$$\frac{1}{1+e^{-b}} - \frac{e^{-b}(1-e^{-b})\,kc}{2(1+e^{-b})^2(k-2)}. \tag{5·5}$$

Table 2 shows for $r = 2$, $k = 20$ and a selection of b, c the exact natural reduction and the closeness of approximation (5·3) to the exact value and its superiority over the modal reduction. All methods are fairly insensitive to changes in k provided k is not too small, which is unlikely in any practical situation.

6. Relation to other models

Aitchison & Kay (1975) have set out for simple data a set of axioms for statistical diagnosis which leads to the use of predictive diagnosis in the sense of Geisser (1964) and Dunsmore (1966). They envisage the generation of a case record (t, x) in a commonly accepted clinical way as a two-stage process in which the type t is first determined by a probabilistic mechan-

ism $p(t|\psi)$; the ψ parameter is the vector of incidence rates of the pure types. Given the type t, the feature vector x is then determined by a conditional probability mechanism $p(x|t,\theta)$, where θ denotes the collection of parameters associated with the relationships of feature to types and so may be called the structural parameter. Thus $p(t,x|\psi,\theta) = p(t|\psi)\,p(x|t,\theta)$. On the basis of a set z of case records $z_1 = (t_1, x_1), \ldots, z_n = (t_n, x_n)$ and the assumption that the prior distribution $p(\psi,\theta)$ for (ψ,θ) factorizes as $p(\psi)\,p(\theta)$, the application of the laws of conditional probability lead to a composite diagnosis $u = (u_1, \ldots, u_r)$ for a new patient with feature vector x:

$$u_t \propto p(t|z)\,p(x|t,z). \tag{6·1}$$

In (6·1), $p(t|z)$ can be interpreted as the composite diagnosis for the new patient based on past cases alone, and $p(x|t,z)$ is the predictive density function for a feature vector of a case of type t evaluated at the actual observed feature vector x of the new patient.

The model represented in §4 is capable of a similar interpretation because we could write $p(v,x|\mu,\tau)$ as a product of a marginal distribution for v and an associated conditional distribution for x. We can express the density function for (v,x) in the form $p(v|\psi)\,p(x|v,\theta)$, where $p(v|\psi)$ is $\mathrm{No}_{r-1}(\lambda,\rho)$ and $p(x|v,\theta)$ is $\mathrm{No}_f(\beta+\gamma v,\delta)$. Parameters (ψ,θ) or $(\lambda,\rho,\beta,\gamma,\delta)$ are related to the parameters (μ,τ) of the model of §4 through the one-to-one relationship $\lambda = \mu_v$, $\rho = \tau_{vv} - \tau_{vx}\tau_{xx}^{-1}\tau_{xv}$, $\beta = \mu_x + \tau_{xx}^{-1}\tau_{xv}\mu_v$, $\gamma = -\tau_{xx}^{-1}\tau_{xv}$ and $\delta = \tau_{xx}$. We can then consider a prior distribution on (ψ,θ) as before leading to a predictive form of diagnosis analogous to (6·2) with

$$p(v|x,z) \propto p(v|z)\,p(x|v,z). \tag{6·2}$$

Thus our model effectively contains as a special case the natural extension of predictive diagnosis described by Aitchison & Kay (1975). For full agreement between (4·1) and (6·2) all that is necessary is the conformability of the prior distributions on (μ,τ) and (ψ,θ).

The form (6·2) allows the disengagement of v from the multinormal requirement of Assumption 1. If the basic composite diagnoses $v_1, \ldots, v_n$ do not conform to multinormality we can retain multinormality of the conditional density function $p(x|v,\theta)$ but adopt an appropriate alternative form for $p(v|\psi)$. The marginal distribution of x is then not necessarily multinormal so that with this approach polymodality of the overall feature vector distribution can be accommodated.

With this approach we can also investigate a desirable limiting property of our diagnostic system. It is a condition of our diagnostic system that a composite diagnosis is a positive probability distribution on T. Hence our system does not directly contain as a special case a diagnostic system based on simple data, that is where $u_t = 1$ for some $t \in T$. It can, however, be considered as a special limiting case where each $u_i\,(i = 1, \ldots, n)$ in the basic set of data tends to one of the set of certain composite diagnoses $e_1, \ldots, e_r$, where e_t has tth component 1, all other components 0.

The other main way in which simple case records have been regarded as generated is by way of the approach of Cox (1966), Day & Kerridge (1967) and Anderson (1972). Again the generation of a case record (t,x) is envisaged as a two-stage process with the feature vector x being first determined by a probabilistic mechanism $p(x|\xi)$, where ξ may be termed the feature parameter. Given the feature vector x the type t is then determined by a conditional probability mechanism $p(t|x,\eta)$, where η now plays the role of a structural parameter. Thus

$$p(t,x|\xi,\eta) = p(x|\xi)\,p(t|x,\eta).$$

From a statistical diagnostic point of view interest then centres on $p(t|x,\eta)$ and so the use of past records $(t_1,x_1),\ldots,(t_n,x_n)$ really involves the regression of the categorical variable t on the feature vector x in some suitable way, usually by a logistic approach. In this approach there is thus no need to make any strong distributional assumption on x, an aspect which cannot be retained in our extension.

Again the model presented in the previous section is capable of a similar interpretation, by writing

$$p(v,x|\mu,\tau) = p(x|\xi)\,p(v|x,\eta), \tag{6·3}$$

where in our distributional form $p(x|\xi)$ is $\mathrm{No}_f(\lambda,\rho)$ and $p(v|x,\eta)$ is $\mathrm{No}_{r-1}(\alpha+\beta x,\gamma)$. The parameters (ξ,η) or $(\lambda,\rho,\alpha,\beta,\gamma)$ are again related to the parameters (μ,τ) of the model of §4 through the relationship $\lambda=\mu_x$, $\rho=\tau_{xx}-\tau_{xv}\tau_{vv}^{-1}\tau_{vx}$, $\alpha=\mu_v+\tau_{vv}^{-1}\tau_{vx}\mu_x$, $\beta=-\tau_{vv}^{-1}\tau_{vx}$ and $\gamma=\tau_{vv}$.

Since in our model v is related to u, the type probability vector, by a logistic relationship we have thus within this interpretation of our model a generalization of the simple-data logistic model of Cox, Day, Kerridge and Anderson to the composite-data case. One advantage of our extension is the purely technical mathematical point that the composite diagnosis variable v is continuous in R^{r-1} compared with the discrete form of the simple diagnosis t, and so regression problems are very much simpler to handle. A second statistical advantage stems from this first mathematical advantage. The Cox, Day, Kerridge and Anderson approach is estimative in that they simply obtain maximum likelihood estimates $\hat{\eta}(z)$ for η and then quote as composite diagnosis $u_t=p\{t|y,\hat{\eta}(x)\}$, and make no allowance for any sampling variability. The advantage of the predictive method (Aitchison & Kay, 1975) can be easily obtained through the form (6·3), and a suitable conjugate prior distribution on (ξ,η) or simply on $\eta=(\alpha,\beta,\gamma)$. With appropriate conformity of this prior to that of Assumption 3 we can obtain the same result as in (4·1). In short, our model extends the logistic regression model approach from simple to composite diagnosis and also from the doubtfully valid estimative approach to the more realistic predictive approach.

The philosophy of the first marginal-conditional approach discussed in this section is that in the basic data diagnosis is determined by factors outside the feature vector x whereas in the second approach diagnosis in the basic data is effectively determined by the feature vector. Our model is symmetric in (v,x) and so allows any relationship between v and x in the basic data to speak for itself.

7. PRACTICAL CONSIDERATIONS

All the results so far considered have been on the assumption that new case records are arising by exactly the same probabilistic mechanism which produced the case records of the basic set. It is in this sense that we have a natural informative experiment with case records arising from the natural experience or incidence pattern at the diagnostic clinic in question. In some circumstances the cases of the basic set are selected and may not reflect the incidence pattern expected in the future. For example, if one of the types is rare we may have had to seek out more than the natural frequency of cases in order to obtain sufficient information on the structural parameter. There is another way in which an informative experiment may turn out to be designed rather than natural. Although the basic case records may have arisen in natural sequence at a clinic we may realize that there has been some

temporal change in the incidence pattern or that we wish to transfer the diagnostic technique to another clinic for which the incidence pattern is known to be different.

An exact approach to accommodate this difference proceeds as follows. From the previous analysis we can extract the information concerning the variability of feature vector for given composite diagnosis, the predictive density function $p(x|v,z)$ of (6·2). This will apply unchanged by any alteration in incidence pattern. Suppose that the new incidence pattern is described by $p(v)$. Then the distributional diagnosis appropriate to this new incidence pattern is given by $p(v|x,z) \propto p(v)p(x|v,z)$. It is unlikely that $p(v)$ and $p(x|v,z)$ will be so conformable as to allow a simple expression of $p(v|x,z)$ in one of the standard forms, so that direct computation is required.

An alternative approximate but intuitively appealing technique is as follows. Consider the diagnostic distribution.

$$p(v|x,z) = \mathrm{St}_{r-1}[k, b_v + c_{vx}c_{xx}^{-1}(x-b_x), (c_{vv} - c_{vx}c_{xx}^{-1}c_{xv})\{1+k^{-1}Q(x)\}]. \qquad (7\cdot1)$$

Suppose that the new incidence pattern is characterized by a mean B_v and covariance matrix C_{vv}. Since a change in the incidence pattern has no effect on the correlation structure of x with v we can alter (7·1) to accommodate the mean and covariance structure for v in the following way. First, we determine square matrices a_v and A_v such that $a_v a_v' = c_{vv}$ and $A_v A_v' = C_{vv}$ and then replace the mean, $b_v + a_v a_v^{-1} c_{vx} c_{xx}^{-1}(x-b_x)$ by $B_v + A_v a_v^{-1} c_{vx} c_{xx}^{-1}(x-b_x)$ and the first factor of the variance parameter similarly by $A_v\{I - a_v^{-1}c_{vx}c_{xx}^{-1}c_{xv}(a_v^{-1})'\}A_v'$. If the new incidence information is based on about the same number of, or indeed more, case records than the original basic set then the k could remain unaltered. If, however, we are aware that fewer case records were available on which to base the new incidence pattern then there would be an argument for the wisdom of reducing k by the deficit.

We have expressed our results in § 4 and subsequent sections in terms of a general normal-Wishart prior on (μ, τ) as specified in Assumption 3. Often in practice there will be no convenient way of determining suitable parameters b, c, g and h and recourse will have to be made to the adoption of a vague prior. This is not a device of despair since it retains the advantages of predictive diagnosis indicated in § 4. In our own applications of predictive diagnosis for simple data we have used the extreme vague prior, which in the context of composite data diagnosis takes the form $p(\mu,\tau) \propto |\tau|^{-\frac{1}{2}(r+f)}$. For this case (4·4) takes the form

$$p(v|x,z) = \mathrm{St}_{r-1}\Big[n-1, \bar{v} + S_{vx}S_{xx}^{-1}(x-\bar{x}), (1+1/n) \times \frac{1}{n-1}(S_{vv} - S_{vx}S_{xx}^{-1}S_{xv})\{1+(n-1)^{-1}Q(x)\}\Big], \qquad (7\cdot2)$$

where $Q(x) = (x-\bar{x})'\{(1+1/n)\,S_{xx}/(n-1)\}^{-1}(x-\bar{x})$.

The appearance of $Q(x)$ in the variance parameter of the Student form (7·2) has the desirable effect that the larger is $Q(x)$ the less certainty there is in the diagnostic statement, other things such as the mean parameter being equal. Roughly speaking, therefore, the further a feature vector is from the centre of the basic set the more cautiously the system approaches the diagnosis. This is a particularly important precaution for new cases which fall 'outside previous experience' in the sense that their feature vectors are outside the basic data cluster. It is easily shown that such precaution is not exercised by the corresponding estimative method.

8. DISCUSSION

The generalization of § 4 gives insight into the diagnostic process, in particular the hierarchical structure of diagnostic statements and the interrelationships of other diagnostic methods. Further work is clearly necessary:

(i) to produce direct algorithms for the evaluation of the natural reduction integral (5·1) to replace the approximate methods for the cases $r \geqslant 3$;

(ii) to extend the analysis to data which are not multinormal;

(iii) to discover the precise role and usefulness of hyperdistributional diagnoses as an expression of panel opinion;

(iv) to evaluate the practical consequences of the availability of statistical diagnostic techniques for composite data.

The traditional view of disease as an either-or phenomenon, i.e. that the aim of diagnosis is to separate patients into clearly defined types, such as normotensive and hypertensive, has been increasingly under attack in recent years from within the medical profession; see, for example, Pickering (1968, pp. 1–5). The admission of such concepts of degree of disease clearly reinforces the need for statistical diagnostic techniques based on composite data.

The authors wish to thank Professor D. R. Cox for a helpful discussion and, in particular, for the suggestion of the normal approximation method for evaluating the natural reduction of (5·2).

REFERENCES

AITCHISON, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–54.

AITCHISON, J. & DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

AITCHISON, J. & KAY, J. W. (1975). Principles, practice and performance in decision-making in clinical medicine. In *Proc. 1973 NATO Conference on the Role and Effectiveness of Decision Theories in Practice*, Eds. K. C. Bowen and D. J. White. London: English Universities Press.

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

COX, D. R. (1966). Some procedures associated with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman*, Ed. F. N. David, pp. 57–71. New York: Wiley.

COX, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

DAY, N. E. & KERRIDGE, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–23.

DUNSMORE, I. R. (1966). A Bayesian approach to classification. *J. R. Statist. Soc.* B **28**, 568–77.

FERRISS, J. B., BROWN, J. J., FRASER, R., KAY, A. W., LEVER, A. F., NEVILLE, A. M., O'MUIRCHEARTAIGH, I. G., ROBERTSON, J. I. S. & SYMINGTON, T. (1970). Hypertension with aldosterone excess and low plasma-renin: preoperative diagnosis between patients with and without adrenocortical tumour. *Lancet* **2**, 995–1000.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–88.

GEISSER, S. (1964). Posterior odds for multivariate normal classifications. *J. R. Statist. Soc.* B **26**, 69–76.

LACHENBRUCH, P. A. & MICKEY, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1–10.

PICKERING, G. (1968). *High Blood Pressure*. London: Churchill.

SIEGEL, C. L. (1935). Über die analytische Theorie der quadratischen Formen. *Ann. Math.* **36**, 527–606.

SMITH, C. A. B. (1947). Some examples of discrimination. *Ann. Eugen.* **13**, 272–82.

[*Received February* 1975. *Revised July* 1975]

AITCHISON, J. (1977)

A calibration problem in statistical diagnosis:
The system transfer problem

Reprinted from *Biometrika* 64, 461-72

Biometrika (1977), 64, 3, *pp.* 461–72
With 1 text-figure
Printed in Great Britain

A calibration problem in statistical diagnosis: The system transfer problem

BY J. AITCHISON
Department of Statistics, University of Hong Kong

SUMMARY

A statistical diagnostic system devised for a particular clinic may require appreciable modification when either the method of measurement of any of the diagnostic features changes within the clinic, or the system is to be applied to another clinic where different methods of measurement are used. In either circumstance the transfer can be effectively made only on the basis of information from a calibration experiment designed to establish the relationship between the different methods of measurement. Even with such calibration information a naive method of calibration commonly adopted is shown to yield conclusions which can be radically different from those of a method taking full account of the calibration unreliability. An application to a particular problem of clinical diagnosis is used to illustrate the analysis.

Some key words: Calibration; Diagnosis; Discriminant analysis; Predictive distribution.

1. INTRODUCTION

The need to develop the statistical techniques of this paper arose directly from two particular and related problems in diagnosis in clinical medicine. For differentiating between two types of Conn's syndrome, a rare disease producing high blood pressure, a statistical diagnostic system, assigning odds on the basis of observations of eight features of a patient's condition, had been developed for a particular clinic A; see Aitchison & Dunsmore (1975) for both the data and an account of the system. After use of the system for some years the clinic has changed its method of measurement of one of the features, plasma concentration of the hormone aldosterone, from a double isotope assay to a less expensive, more efficient radioimmunoassay. Any new case of Conn's syndrome will therefore have this hormone measured by the new method only.

Since the patients forming the diagnostic data base or training set have been discharged or are undergoing treatment which affects the hormone concentration it is impossible to find the radioimmunoassay counterparts of their original double isotope assay measurements in order to construct a new diagnostic system directly from these patients. Fortunately, however, although the concentration of hormone depends on the type of the syndrome the conditional distributions relating one hormone determination to the other are known not to depend on type nor indeed on the fact that the patient has the particular syndrome. We can therefore investigate the possibilities of calibrating from the new to the old by measuring hormone concentration by both methods on portions of blood from a number of portions not in the training set and not necessarily suffering from the disease. Such pairs of observations are available on 72 blood samples. It is tempting to suppose that all that is then necessary is to plot a scatter diagram, to fit some form of regression line and, for a given radioimmunoassay measurement, to read off from the regression line the corresponding 'calibrated' value of double isotope determination, and finally to use this directly in the original statistical diagnostic system devised for clinic A. Unfortunately such a method, though simple, takes no account of

the unreliability of the calibrated value. A main aim of this paper is to analyze the extent to which this naive calibration method may differ from an approach which takes full account of the unreliability.

The second problem arose when requests were received from another clinic B to process data from its patients through the statistical diagnostic system devised for clinic A. Uncritical application of the system to data sent from clinic B on its patients can give rise to serious failures with a subsequent loss of confidence in the system in both clinics. The reason for such failures is almost invariably that the clinics use different methods of determining some or all of the features on which the statistical diagnosis is based. Again the circumstances and the methods in the particular example were such that calibration information could be obtained from blood samples on other patients. Thus statistically these system transfer problems are identical, clinic B being interpreted either as different from clinic A and with different methods of measurement, or as clinic A after it has changed to its new methods of measurement. It is also notionally and notationally convenient in our early discussion to suppose that the methods of measurement of all features differ from clinic A to clinic B. The adjustment required to cover the situation where only some of the methods of measurement differ is discussed in § 6.

The data available for this system transfer problem arise from three sources: first, diagnostic data

$$D_A = \{(t_i, x_i); \quad i = 1, \ldots, n\}$$

from clinic A consisting of known types t_i and known feature vectors x_i of n of its patients forming the diagnostic training set; secondly, calibration data

$$C_{AB} = \{(y_{Aj}, y_{Bj}); \quad j = 1, \ldots, k\}$$

consisting of determinations of feature vectors made by both clinics on k individuals, y_{Aj} referring to measurement by clinic A and y_{Bj} the corresponding measurement by clinic B. Thirdly, for a new case of unknown type t, we have observed the feature vector x_B by the methods of clinic B. The problem is then to model the situation so as to obtain a realistic assessment of the conditional or diagnostic probabilities $p(t|x_B, D_A, C_{AB})$, the plausibilities we attach to the possible disease types for this new case on the basis of the diagnostic and calibration data and of the patient's own feature vector.

This modelling is set out in § 3, but first we require in § 2 some comments on the nature of the separate problems of diagnosis and calibration. In the remaining sections the consequences of the model are worked out in §§ 4 and 5, leading to the resolution of the particular motivating problem in § 6.

2. STATISTICAL DIAGNOSIS AND CALIBRATION

We adopt the view that the statistician's role in diagnosis is, for each referred patient, to present the clinician with a realistic assessment of probabilities for each of the disease types within the set T of possible types. Since the problem of adjusting an assessment of these diagnostic probabilities based on one incidence pattern of types to take account of a different incidence pattern is statistically trivial, requiring only proportional adjustments to the assessed probabilities, we can for convenience assume equal incidence rates for each type.

For a patient in clinic A of unknown type t and known feature vector x_A the diagnostic problem is to assess the values of $p(t|x_A, D_A)$. Since with equal incidence rates

$$p(t|x_A, D_A) \propto p(x_A|t, D_A)$$

the statistical problem can be considered as that of placing realistic values on $p(x_A|t, D_A)$ for

each of the possible types. A common procedure is to suppose that the true density functions of feature vector for given type belong to some parametric class, say $p(x_A|t,\theta)$, where the indexing parameter $\theta \in \Theta$. Within this framework there are two main methods of assessment, given the training set D_A. The estimative method simply uses D_A to obtain an efficient estimate, say $\hat{\theta}(D_A)$, of θ and then substitutes estimate for parameter in the parametric form, obtaining

$$p\{x_A|t,\theta = \hat{\theta}(D_A)\}.$$

The predictive method first converts a prior distribution $p(\theta)$ on Θ through Bayes's formula to a posterior distribution $p(\theta|D_A)$, which is then used as a mixture weighting of the parametric forms to obtain the predictive density function

$$p(x_A|t,D_A) = \int_\Theta p(x_A|t,\theta)\,p(\theta|D_A)\,d\theta. \tag{2·1}$$

Aitchison, Habbema & Kay (1977) point out the considerable differences between these two methods when the training set D_A is limited in size, and demonstrate the greater realism provided by the predictive method. We therefore adopt the predictive method in our diagnostic assessments and will indeed find that the differences between the two diagnostic methods can be accentuated by the presence of a calibration problem, and in particular by the use of the naive calibration method.

We shall find in our later analysis that calibration in the context of the problem of § 1 involves us in the assessment of the conditional density functions $p(x_B|x_A,C_{AB})$. Statistically the calibration problem is formally similar to that of statistical diagnosis, except that a continuous set X_B now replaces the finite set T. How we carry out the assessment of the conditional density function depends on the design of the calibration experiment, whether the (y_{Aj},y_{Bj}) arise in a natural bivariate way or whether one set of measurements, say the y_{Aj}, has been selected or controlled. Careful model formulation to conform with this design is therefore necessary to determine for example whether we arrive at $p(x_B|x_A,C_{AB})$ directly by introducing a parametric form $p(x_B|x_A,\delta)$ with $\delta \in \Delta$ and then setting

$$p(x_B|x_A,C_{AB}) = \int_\Delta p(x_B|x_A,\delta)\,p(\delta|C_{AB})\,d\delta, \tag{2·2}$$

or through Bayes's formula

$$p(x_B|x_A,C_{AB}) \propto p(x_B|C_{AB})\,p(x_A|x_B,C_{AB})$$

and by introducing a parametric form $p(x_A|x_B,\delta)$. For the motivating problem we shall find that (2·2) is the appropriate version and is in the form of predictive calibration. Estimative calibration, by analogy with estimative diagnosis, would replace δ by an efficient estimate $\hat{\delta}(C_{AB})$ and use

$$p\{x_B|x_A,\delta = \hat{\delta}(C_{AB})\} \tag{2·3}$$

instead of (2·2). Naive calibration goes a step further along the road of throwing aside considerations of reliability of inference assessments by replacing the estimative distribution (2·3) by a point estimate. In our criticism of naive calibration we shall again prefer predictive rather than estimative calibration as our main standard of comparison.

3. A STATISTICAL MODEL FOR CALIBRATED DIAGNOSIS

To model the clinical situation described in § 1 a first step is to postulate a sensible probabilistic mechanism for the generation of a complete record (t, x_A, x_B) for an individual case, where t is the type of the case and x_A and x_B are the feature vectors of the case as measured in clinics A and B, respectively. Postulating such a model does not imply either that complete records must be available in the data set or that we contemplate the determination of complete records for new cases. This basic model provides a conceptual link between type and the two clinics' methods of measurement. It simply consists of formulating the joint distribution of (t, x_A, x_B) in terms of conditional distributions together with assumptions concerning separability of parameters and independence of data sets. To provide an operational model we must pay attention to a number of interrelated factors.

(i) The model should allow an expression of our understanding of the type–feature relationship in clinic A.

(ii) It should allow an adequate description of the nature of the calibration experiment.

(iii) It must allow a derivation of a likelihood function for the observed data C_{AB}, D_A and x_B.

(iv) It must allow the fulfilment of the purpose of the investigation, in our case, the assessment of the probabilities $p(t|x_B, C_{AB}, D_A)$.

For convenience of reference we first state six assumptions of a model for the system transfer problem, discuss their relevance to the particular problem of § 1, and then deduce from the assumptions the appropriate method of calibrated diagnostic assessment.

Assumption 1. Any case has associated with it a unique type belonging to a finite set T of possible types, and possesses a feature vector belonging to a set X_A or X_B of possible feature vectors depending on whether the features are observed in clinic A or clinic B.

Assumption 2. The model is parametric with parameter set Ω, so that the model corresponding to $\omega \in \Omega$ is specified by the density function

$$p(t, x_A, x_B|\omega) \quad (t \in T, x_A \in X_A, x_B \in X_B).$$

Assumption 3. In the conditional specification

$$p(t|\omega)\, p(x_A|t, \omega)\, p(x_B|t, x_A, \omega)$$

of $p(t, x_A, x_B|\omega)$ the parameter ω and the parameter set Ω can be factorized into $\omega = (\psi, \theta, \delta)$ and $\Omega = \Psi \times \Theta \times \Delta$ in such a way that they are separable:

$$p(t, x_A, x_B|\omega) = p(t|\psi)\, p(x_A|t, \theta)\, p(x_B|t, x_A, \delta).$$

Assumption 4. We have $p(x_B|t, x_A, \delta) = p(x_B|x_A, \delta)$.

Assumption 5. Given ψ, θ, δ and the type t of the new case from clinic B the data sets C_{AB}, D_A and x_B are independent.

Assumption 6. There is prior independence in that $p(\psi, \theta, \delta) = p(\psi)\, p(\theta)\, p(\delta)$.

In the clinical setting our first assumption is standard in statistical diagnosis, asserting that the disease types have been defined so as to be mutually exclusive and exhaustive, and that symptoms, signs and the results of diagnostic tests constituting a feature vector can be obtained for each patient. Our second assumption merely acknowledges that we are adopting a parametric model.

In the third assumption the separation of ψ and θ in the first two factors follows the usual assumption adopted in diagnostic models (Aitchison & Dunsmore, 1975, Chapter 11) which recognize ψ as an incidence parameter and θ as a structural parameter, in the sense that it

reflects the possible dependence of the feature variability on the disease type. The separation of δ from θ and ψ is also often reasonable. Its only implication is that if we know t and x_A then the distribution of x_B can be indexed by a separate parameter. For example, in the case where $p(x_A, x_B|t, \omega)$ is multivariate normal, this can always be achieved by θ referring to marginal mean vectors and covariance matrices and δ to the familiar and separable parameters of the regression distributions of x_B on x_A.

The fourth assumption is one which will require careful scrutiny in any particular application. It asserts that the calibration relationship does not depend on type. In the case of Conn's syndrome, although a patient's plasma concentration of a substance such as aldosterone certainly depends on type the relationship of radioimmunoassay determination to double isotope assay determination is one which holds irrespective of type or indeed irrespective of whether the blood sample comes from a patient with Conn's syndrome. All that was necessary therefore was to select a range of blood samples to meet the obvious calibration requirements that they cover the set of future values likely to arise, and to make determinations by both methods on portions of these samples. If this assumption does not hold then we may be in great difficulty. For example, if x_A were the concentration, measured in mg/dl, of a hormone in urine and x_B were the urinary excretion rate, measured in mg/day, and if patients with disease type 1 had a tendency to retain body fluid compared with patients of disease type 2, then the relationship of x_B to x_A would clearly depend on type. To assess the density functions $p(x_B|x_A, t, \delta)$ we would require to be able to observe the features in both clinics for patients of each possible type, and this may be physically impossible. It is then only realistic to admit that no reliable system transfer can be achieved and that the only course open is to start building up a new diagnostic training set within clinic B.

Since the three data sets, C_{AB}, D_A and the feature vector x_B of the new case from clinic B, are associated with completely different sets of individuals in our practical situations the fifth assumption automatically applies.

Our final assumption is common in such Bayesian formulations (Aitchison & Dunsmore, 1975, Chapters 10, 11) and asserts that any prior information concerning ψ, θ and δ arises from independent sources. The vague priors that we adopt in practice to ensure no overstatement of odds satisfy this assumption. Its great merit is that of mathematical tractability by ensuring, with Assumptions 3 and 5, that the posterior distribution for these parameters separates in exactly the same way as the prior.

The consequences of these assumptions can easily be worked through in terms of the likelihood for t, ψ, θ and δ given x_B, D_A and C_{AB}. We omit the tedious details so as to pinpoint the main steps in the derivation. Remembering our remark about the trivial nature of incidence rate adjustment and also noting the separability of ψ implied by Assumptions 3 and 6 we can effectively ignore ψ in the argument. Then the essential step in forming the likelihood $L(t, \theta, \delta|x_B, D_A, C_{AB})$ is to note that

$$\begin{aligned} p(x_B|t, \theta, \delta) &= \int_{X_A} p(x_B|x_A, t, \theta, \delta)\, p(x_A|t, \theta, \delta)\, dx_A \\ &= \int_{X_A} p(x_B|x_A, \delta)\, p(x_A|t, \theta)\, dx_A \end{aligned} \tag{3·1}$$

by Assumptions 3 and 4. The likelihood can thus, by Assumption 5, be expressed in the form $p(x_B|t, \theta, \delta)\, p(D_A|\theta)\, p(C_{AB}|\delta)$. Assumption 6 about the factorization of the prior distribution then ensures that the posterior distribution can be expressed in the form

$$p(t, \theta, \delta|x_B, D_A, C_{AB}) \propto p(x_B|t, \theta, \delta)\, p(\theta|D_A)\, p(\delta|C_{AB}),$$

invoking the equal incidence assumption. Integration with respect to θ and δ then gives

$$p(t|x_B, D_A, C_{AB}) \propto \int_\Theta \int_\Delta p(x_B|t,\theta,\delta)\, p(\theta|D_A)\, p(\delta|C_{AB})\, d\theta\, d\delta$$

$$\propto \int_{X_A} p(x_B|x_A, C_{AB})\, p(x_A|t, D_A)\, dx_A \qquad (3{\cdot}2)$$

by (3·1), where

$$p(x_B|x_A, C_{AB}) = \int_\Delta p(x_B|x_A,\delta)\, p(\delta|C_{AB})\, d\delta, \qquad (3{\cdot}3)$$

$$p(x_A|t, D_A) = \int_\Theta p(x_A|t,\theta)\, p(\theta|D_A)\, d\theta \qquad (3{\cdot}4)$$

are the predictive calibrative and diagnostic distributions referred to at (2·2) and (2·1).

The mathematical problem involved in calibrated diagnosis is thus the evaluation of the convolution-type integral (3·2), and this is seldom standard since the calibrative distribution $p(x_B|x_A, C_{AB})$ may involve x_A in a complicated way. We return to this problem in § 5 after consideration of an idealized form of the problem which allows us an insight into its nature without the encumbrance of the inferences involved in the construction of the predictive distributions (3·3) and (3·4).

4. The case of perfect information on calibration and diagnosis

Even if we know exactly all the conditional distributions involved in Assumption 3 the calibration problem remains. To simplify the notation we drop the dependence on ψ, θ and γ in these known distributions, and again work with equal incidence rates. Suppose that there are just two disease types 1 and 2, that the features measurable in clinics A and B are each one-dimensional, and that the clinic A feature distribution $p(x_A|t)$ for given type t is of known form, normally distributed with mean μ_t and variance σ^2. Note the common variance assumption as an additional simplification here. Moreover for a given measurement x_A in clinic A suppose that the conditional distribution $p(x_B|x_A, t)$ of the corresponding measurement x_B in clinic B has a normal linear regression model independent of disease type:

$$p(x_B|x_A, t) = N(\alpha + \beta x_A, \gamma^2).$$

To arrive at a diagnostic assessment $p(t|x_B)$ for a patient, whose feature measurement x_B has been recorded in clinic B only, we have, from (3·2),

$$p(t|x_B) \propto p(x_B|t) = \int_{X_A} p(x_B|x_A, t)\, p(x_A|t)\, dx_A = \phi(x_B|\alpha + \beta\mu_t, \gamma^2 + \beta^2\sigma^2),$$

where $\phi(x_B|\mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ density function. This assessment involves only standard conditional probability arguments with known distributions and hence is exact.

The naive calibration method described in § 1 uses a calibrated point estimate

$$\hat{x}_A = (x_B - \alpha)/\beta$$

as the clinic A counterpart of x_B, and so arrives at a diagnostic assessment

Then $$q(t|x_B) \propto p(x_A = \hat{x}_A|t) = \phi(x_B|\alpha + \beta\mu_t, \beta^2\sigma^2).$$

$$\log\frac{q(1|x_B)}{q(2|x_B)} = \left(\frac{\mu_1 - \mu_2}{\sigma}\right)\left\{\frac{\hat{x}_A - \tfrac{1}{2}(\mu_1 + \mu_2)}{\sigma}\right\}, \qquad (4{\cdot}1$$

$$\frac{p(1|x_B)}{p(2|x_B)} = \left\{\frac{q(1|x_B)}{q(2|x_B)}\right\}^{\lambda/(1+\lambda)}, \qquad (4{\cdot}2)$$

where $\lambda = (\beta\sigma/\gamma)^2$.

From (4·2) we see that the naive calibration method gives more extravagant odds since $p(1|x_B)/p(2|x_B)$ is always further away than $q(1|x_B)/q(2|x_B)$ from the equiprobable ratio 1. This is an intuitively obvious result since the naive calibration method is making no allowance for the unreliability of the calibrate $\hat{x}_A$. The form of (4·1) shows that the extent of the extravagance depends on the combination $\beta\sigma/\gamma$; the larger this quantity is the nearer the naive calibration method is to the exact method. The ratio β/γ is a standard measure of the effectiveness of a calibrative system; the larger this ratio the more reliable the calibration. The reason for the involvement of the factor σ is simply that, for given μ_1 and μ_2, the larger the value σ takes, the less effective is the diagnostic system and consequently the less there is to lose in using the naive calibration method.

There is a multivariate analogue of the preceding analysis, best expressed in terms of a form easily derivable from (4·1) and (4·2):

$$\log\frac{q(1|x_B)}{q(2|x_B)}-\log\frac{p(1|x_B)}{p(2|x_B)}=\left(\frac{\mu_1-\mu_2}{\sigma}\right)\left(\frac{1}{1+\lambda}\right)\left\{\frac{\hat{x}_A-\frac{1}{2}(\mu_1+\mu_2)}{\sigma}\right\}. \tag{4·3}$$

The first factor of (4·3) is the standardized distance between the two means, the second factor a measure of calibration unreliability, and the third factor the standardized distance of the naive calibrate from the centre of the feature distributions. We retain the notation of the univariate case and let both x_A and x_B be d-dimensional, with

$$p(x_A|t)=N_d(\mu_t,\Sigma),\quad p(x_B|x_A,t)=N_d(\alpha+Bx_A,\Gamma),$$

so that

$$p(x_B|t)=N_d(\alpha+B\mu_t,\Gamma+B\Sigma B').$$

For naive calibration we assume that B is nonsingular so that $\hat{x}_A=B^{-1}(x_B-\alpha)$ and

$$p(x_A=\hat{x}_A|t)\propto\phi_d(x_B|\alpha+B\mu_t,B\Sigma B').$$

The difference between the naive and exact log odds is then

$$\begin{aligned}\log\frac{q(1|x_B)}{q(2|x_B)}-\log\frac{p(1|x_B)}{p(2|x_B)}&=(\mu_1-\mu_2)'\{\Sigma^{-1}-B'(\Gamma+B\Sigma B')^{-1}B\}\{\hat{x}_A-\tfrac{1}{2}(\mu_1+\mu_2)\}\\&=\{T^{-1}(\mu_1-\mu_2)\}'(I+T'B'\Gamma^{-1}BT)^{-1}[T^{-1}\{\hat{x}_A-\tfrac{1}{2}(\mu_1+\mu_2)\}],\end{aligned} \tag{4·4}$$

where $TT'=\Sigma$. The expression (4·4) again has three factors. In these T^{-1} plays the same standardization role with respect to the covariance structure Σ as σ^{-1} plays with respect to the variance σ^2. The first factor is thus the standardized vector displacement between the two means, with squared length equal to the Mahalanobis measure $(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$. The third factor is again the standardized vector displacement of the naive calibrate from the centre of the feature vector distributions. The middle factor is associated with the unreliability of the calibration, the matrix $\Lambda=T'B'\Gamma^{-1}BT$ replacing the univariate λ of (4·3).

We can obtain some insight into the possible consequences of having to calibrate by considering the special case where for given type the features are independent so that $\Sigma=\mathrm{diag}\,(\sigma_i^2)$, and where each feature of clinic A is calibrated independently against the corresponding feature in clinic B, so that $B=\mathrm{diag}\,(\beta_i)$, $\Gamma=\mathrm{diag}\,(\gamma_i^2)$. Then

$$(I+T'B'\Gamma^{-1}BT)^{-1}=\mathrm{diag}\{(1+\lambda_i)^{-1}\},$$

where $\lambda_i=(\beta_i\sigma_i/\gamma_i)^2$, and (4·4) becomes

$$\sum_{i=1}^{d}\left(\frac{\mu_{1i}-\mu_{2i}}{\sigma_i}\right)\left(\frac{1}{1+\lambda_i}\right)\left\{\frac{\hat{x}_{Ai}-\frac{1}{2}(\mu_{1i}+\mu_{2i})}{\sigma_i}\right\}.$$

Each term in this sum is of the same form as (4·3) and it is obvious that separately moderate calibrative distortions can accumulate to give a pronounced effect. The extent of the difference between the naive and the exact assessments can be illustrated by two simple examples.

Examples. For $d = 2$ and

$$\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & -0{\cdot}7 \\ -0{\cdot}7 & 1 \end{bmatrix}, \quad B = I_2, \quad \Gamma = \text{diag}\,(0{\cdot}16, 0{\cdot}16),$$

the naive odds for $x_B = \mu_1$ are 28 to 1 on type 1, whereas the exact odds are only 9 to 1. For $d = 3$ and

$$\mu_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & -0{\cdot}7 & 0{\cdot}8 \\ -0{\cdot}7 & 1 & -0{\cdot}9 \\ 0{\cdot}8 & -0{\cdot}9 & 1 \end{bmatrix}, \quad B = I_3, \quad \Gamma = 0{\cdot}25 I_3,$$

the naive odds for $x_B = \mu_1$ are 22 026 to 1 on type 1, whereas the corresponding exact odds are only 25 to 1.

Although illustrative examples cannot establish that such large differences will arise in practice they do suggest that any uncritical application of naive calibration in diagnostic problems should be viewed with some suspicion. It is worth emphasizing here also that the simplicity of the relationships between naive and exact odds depends crucially on the assumption of equality of the covariance structures of the feature distributions for given types. When the covariance structures differ, such as in the real application of § 6, it is difficult to provide any simple general statement of the nature of the relationship. The analysis of the present section must nevertheless prepare us to anticipate differences, and possibly substantial differences, between the odds assigned by the two methods in this more general situation.

We must now consider whether these differences are aggravated or not when the diagnostic and calibrative distributions have to be inferred from data such as D_A and C_{AB}.

5. CALIBRATED DIAGNOSIS FOR THE NORMAL CASE

Suppose that $p(x_A|t, \theta)$ is $N(\mu_t, \sigma^2)$ and $p(x_B|x_A, \delta)$ is $N(\alpha + \beta x_A, \gamma^2)$, and that we adopt the customary vague priors for the parameters and the notation of Aitchison & Dunsmore (1975, Chapter 2); in particular we say that a random variable u follows a generalized Student distribution $\text{St}(k, b, c)$ if $(u-b)/\sqrt{c}$ follows a standard t distribution with k degrees of freedom. Then we have the predictive calibrative density function

$$p(x_B|x_A, C_{AB}) = \text{St}\,[k-2, a+bx_A, c^2\{1+1/k+(x_A-\bar{y}_A)^2/S_A\}],$$

where a, b are the usual regression estimates, c^2 the residual mean square, $\bar{y}_A$ and S_A the mean and corrected sum of squares of the y_{Aj}; and as a basis of predictive diagnosis

$$p(x_A|t, D_A) = \text{St}\{n-1, m_t, s^2(1+n_t^{-1})\},$$

where n_t is the number of x_{Ai} of type t, m_t their mean, and s^2 the usual pooled sample variance. In contrast the effect of applying naive calibration and estimative diagnosis is to replace $p(x_B|t, D_A, C_{AB})$ by a normal distribution $N(a+bm_t, b^2s^2)$.

The integral (3·2) cannot be evaluated directly in terms of known functions but we can obtain a useful approximation for comparison purposes by first determining the mean and variance associated with (3·2) by the usual iterated expectation method and then using a normal approximation with this computed mean and variance. This results in a normal

distribution with the same mean $a+bm_t$ as the naive-calibration estimative-diagnostic method but the variance is

$$\frac{n-2}{n-4}(1+n_t^{-1})\,b^2s^2+\frac{k-2}{k-4}c^2\left\{1+k^{-1}+\frac{(m_t-y_A)^2+s^2(n-2)\,(1+n_t^{-1})/(n-4)}{S_A}\right\} \tag{5·1}$$

instead of the b^2s^2 of the naive-calibration estimative-diagnostic method. It is clear that the source of the difference between the naive and exact methods in the case of perfect information, namely the difference between the variances $\beta^2\sigma^2$ and $\beta^2\sigma^2+\gamma^2$ is now considerably increased because of the additional factors in (5·1).

This increase persists in a d-dimensional multivariate setting with r disease types, where the density function (3·2) is approximated by a normal density function with mean $a+Bm_t$ and covariance matrix

$$\left(\frac{n-r}{n-r-d-1}\right)(1+n_t^{-1})\,BSB'+\left(\frac{k-d-1}{k-2d-2}\right)$$
$$\times G\left\{1+k^{-1}+(m_t-\bar{y}_A)\,S_A^{-1}(m_t-\bar{y}_A)+\left(\frac{n-r}{n-r-d-1}\right)(1+n_t^{-1})\,\mathrm{tr}\,(S_A^{-1}S)\right\}, \tag{5·2}$$

where a, B are the usual regression estimates, G is the residual covariance matrix, $\bar{y}_A$ and S_A the mean vector and corrected sum of squares matrix of the y_{Aj}, m_t the mean vector of the n_t feature vectors x_{Ai} which are of type t, and S the usual pooled covariance matrix. The fact that appreciable differences could exist even in the case of perfect information strongly suggests that care must be taken in any real application.

Before we consider such an application we must consider the case of partial calibration.

6. Partial calibration: An application

In many diagnostic transfer problems there will be a number of features, such as age, for which we may safely assume that observation or measurement will lead to the same value in both clinics. We can thus partition x_A into two subvectors $(x_A^{(1)}, x_A^{(2)})$, where only $x_A^{(2)}$ requires calibration. Then C_{AB} need contain calibrative information only on the subvectors $y_{Aj}^{(2)}$, $y_{Bj}^{(2)}$ while D_A still contains information on the whole vector x_{Ai}. Then if x_B is the complete feature vector of a new patient observed in clinic B we have as the basis of our diagnostic assessment the counterpart of (3·2):

$$p(x_B|t,C_{AB},D_A)$$
$$=p(x_A^{(1)}=x_B^{(1)}|t,D_A)\int_{X_A^{(2)}}p(x_B^{(2)}|x_A^{(2)},C_{AB})\,p(x_A^{(2)}|t,x_A^{(1)}=x_B^{(1)},D_A)\,dx_A^{(2)}, \tag{6·1}$$

where the second factor of the integrand with its extra conditioning on $x_A^{(1)}$ is easily derived from the previous $p(x_A|t,D_A)$.

We can now illustrate in the real medical setting of § 1 the consequences of neglecting the full statistical problem of calibrating for diagnosis. Clinic B is clinic A after it has switched from double isotope assay determination $x_A^{(2)}$ to radioimmunoassay determination $x_B^{(2)}$ of plasma concentration of aldosterone, the techniques of determining the other seven features $x_A^{(1)}$ remaining the same, so that $x_B^{(1)}=x_A^{(1)}$. For multivariate normality considerations we work throughout in terms of transformed data, the natural logarithms of the feature observations. The calibrative experiment with a random selection of 72 blood plasma samples provides the

set C_{AB} of data from which we determine along the predictive lines set out by Aitchison & Dunsmore (1975, § 2·5):

$$p(x_B^{(2)}|x_A^{(1)}, C_{AB}) = \mathrm{St}_1\{70, 0{\cdot}647 + 0{\cdot}749x_A^{(1)}, 0{\cdot}187 + 0{\cdot}00194\,(x_A - 2{\cdot}70)^2\}.$$

The other factor in the integrand of (6·1) can be easily obtained for a particular case. To

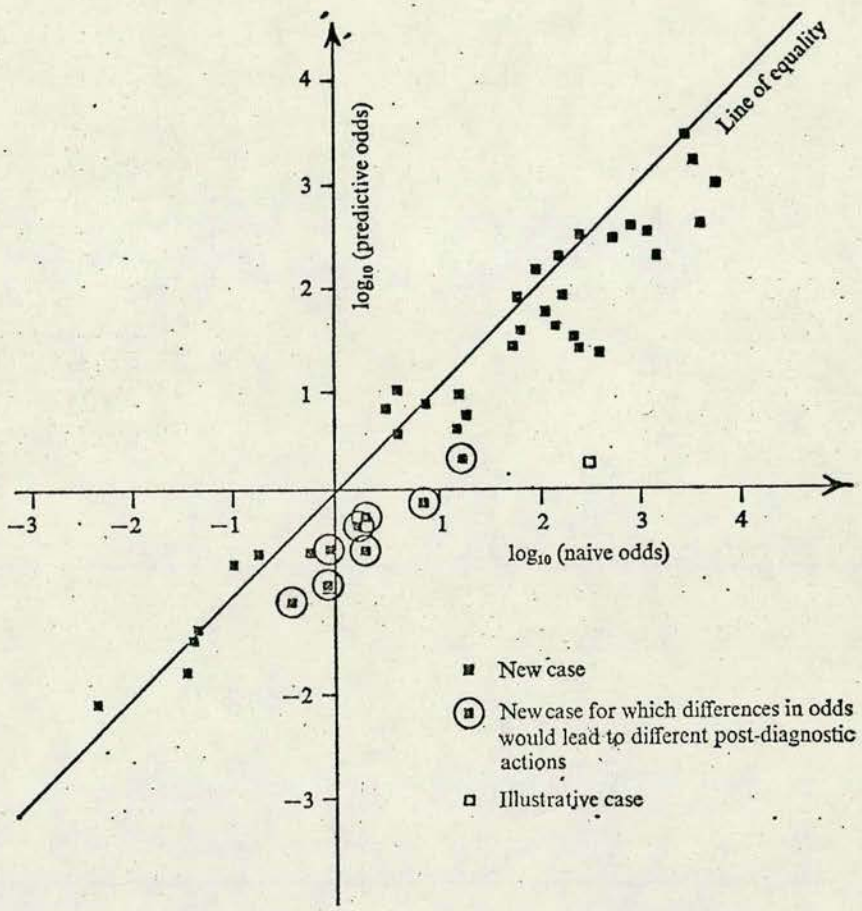

Fig. 1. Comparison for 43 new cases of Conn's syndrome of the predictive diagnostic odds as determined by naive and by predictive calibration.

illustrate the calibration problem we consider the new undiagnosed case of Table 1·6 of Aitchison & Dunsmore (1975) for which

$$p(x_A^{(2)}|t = 1, x_A^{(1)}, D_A) = \mathrm{St}_1(19, 3{\cdot}42, 0{\cdot}483), \tag{6·2}$$

$$p(x_A^{(2)}|t = 2, x_A^{(1)}, D_A) = \mathrm{St}_1(10, 2{\cdot}74, 0{\cdot}0583). \tag{6·3}$$

The theoretical comparison of § 5 has been most conveniently expressed in terms of the contrast between the naive-calibration estimative-diagnosis method and the predictive-calibration predictive-diagnosis method. To emphasize that it is the naive attitude to the calibration problem which can be largely responsible for these differences we compare in our real diagnostic problem the naive and predictive calibration methods within the framework of predictive diagnosis. For both methods the first factor $p(x_A^{(1)} = x_B^{(1)}|t, D_A)$ of (6·1) takes the same seven-dimensional Student form, so that the difference in the methods lies in the treat-

ment of the second factor. The naive calibration method replaces the second factor of (6·1) by

$$p(x_A^{(2)} = \hat{x}_A^{(2)} | x_A^{(1)}, t, D_A),$$

where $\hat{x}_A^{(2)} = (x_B^{(2)} - 0{\cdot}647)/0{\cdot}749$ is the naive calibrate corresponding to $x_B^{(2)}$.

For the particular case under discussion this naive method leads to odds of 310 to 1 in favour of adenoma. The predictive calibration method requires the numerical evaluation of the integrals associated with the right side of (6·1). This can be easily achieved by numerical integration even on a simple desk calculator such as the Hewlett–Packard 9810, and for this particular case leads to odds of only 2 to 1 on adenoma. The difference between these two sets of odds would lead to quite different treatments for the patient. Statisticians must therefore be careful they do not, by using naive calibration, read too much into interclinic data but take full account of the extent of the unreliability of the calibration process.

This already published case has been used simply as a convenient and dramatic means of illustrating the contrast in the naive and predictive calibrative methods of resolving the system transfer problem in diagnosis. For it the assay method was in fact identical to that used in the training set so that the above analysis is only a pointer to what might happen. In the analysis of 43 unpublished new cases where the assay methods differed from that of the training set, the odds assigned by the naive and predictive calibrative methods are shown in Fig. 1. We note here that the 'variance parameters' of the conditional feature distributions (6·2) and (6·3) associated with the published case are substantially different. The same is true for the conditional feature distributions of these 43 unpublished cases. For these cases therefore we recall the comment in § 4 that we cannot anticipate any definite relationship pattern between naive and predictive odds. There are, however, obviously appreciable departures from the line of equal odds assessments. Where both naive and predictive odds both exceed, or are both below, unity by a substantial factor such differences would have little effect on the treatment of the case. Although there is no case displaying as extreme a difference as the illustrative case, there do remain eight cases, identified by open circles in Fig. 1, where the naive and predictive calibrative methods are sufficiently different to lead to important differences in the assessment of the next step in the clinical management.

7. Discussion

Although we have presented the system transfer problem in a medical setting it is clear that for any problem of discriminant analysis, where different techniques of measurement may have been used by different investigators or laboratories, the effects of calibration unreliability should be carefully studied. That such effects can be of practical significance has been amply shown by the clinical cases of this paper. Particular aspects needing further work are:

(i) investigation of the extent to which the approximations suggested in (5·1) and (5·2) are adequate alternatives to the numerical evaluation of the integral (3·2);

(ii) modelling the situation where clinic B selects the cases for the calibration experiment, so that the conditioning adopted in Assumption 3 is inappropriate;

(iii) combining limited diagnostic data from several clinics, by the use of calibrative information, to produce an effective diagnostic system for each clinic;

(iv) removing the limitation of the ideas illustrated here to the multinormal feature distributions by considering more general feature distributions involving binary, discrete and continuous measurements. In this much more difficult problem a useful approach may be to assess such distributions through the general kernel technique proposed by Aitchison & Aitken (1976, § 5).

REFERENCES

AITCHISON, J. & AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* 63, 413–20.

AITCHISON, J. & DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

AITCHISON, J., HABBEMA, J. D. F. & KAY, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statist.* 26, 15–25.

[*Received August* 1976. *Revised April* 1977]

AITCHISON, J. (1979)

A calibration problem in statistical diagnosis:
The clinic amalgamation problem

Reprinted from *Biometrika* 66, 357-66

Biometrika (1979), 66, 2, *pp.* 357–66
Printed in Great Britain

A calibration problem in statistical diagnosis: The clinical amalgamation problem

BY J. AITCHISON
Department of Statistics, University of Hong Kong

SUMMARY

To obtain a sufficiently large diagnostic training set for differential diagnosis it is often necessary to use cases from two or more clinics. Important questions that then arise are whether there are any differences between clinics in their methods of measuring the diagnostic features of cases, to what extent ignoring such differences invalidates any diagnostic system devised for these clinics and in what manner interclinic calibration information may be employed to make efficient use of the complete set of calibrative and diagnostic data. The modelling of such complex diagnostic situations is discussed, problems of statistical methodology arising are resolved, and the methods devised are illustrated by a particular problem of amalgamation for diagnostic purposes of data from different clinics.

Some key words: Calibration; Diagnosis; Discriminant analysis.

1. INTRODUCTION

When diagnostic problems involve two or more clinics in which methods of measurements of diagnostic features differ then important problems of the appropriate form of calibration between clinics arise. For the simplest of these problems, that of system transfer where a diagnostic system devised wholly within one clinic is to be applied to a patient measured in another clinic, Aitchison (1977a) has shown that considerable care in statistical modelling is required if misleading results are to be avoided in practice. In particular, any method which ignores the inherent imprecision of the calibration process may provide assessments of diagnostic probabilities differing substantially from those assigned by methods which fully recognize this aspect of imprecision. In the present paper we develop methods for a more complex calibrative-diagnostic problem in which two or more clinics wish to pool their diagnostic data in order to construct a more reliable diagnostic system than any one clinic could produce by itself. Indeed when the differential diagnosis of a set of rare diseases is involved it may be impossible for any one clinic to obtain enough cases to make construction of a diagnostic system a feasible proposition. When methods of measurement of diagnostic features differ from clinic to clinic or have changed over time within a clinic then we have a calibrative problem of diagnosis which we can conveniently term the clinic amalgamation problem. This problem was posed by Aitchison (1977a) and an indication of how it may be modelled is given in abstract form by Aitchison (1977b). Here we develop these indications and demonstrate the method in simple applications.

2. A CLINIC AMALGAMATION MODEL FOR TWO CLINICS AND TWO TYPES

The nature of the clinic amalgamation problem can be clearly seen from its very simplest form involving only two clinics, labelled 1 and 2 and faced with the problem of differential diagnosis between just two mutually exclusive disease types, say types 1 and 2. This simple

form is appropriate for the application which motivated this paper, and we leave discussion of the extensions to more than two clinics and to more than two disease types until § 6. We suppose that we have two independent diagnostic training sets D_1 and D_2 in the two clinics, with

$$D_i = \{(t_{ij}, x_{ij}); j = 1, \ldots, n_i\} \quad (i = 1, 2),$$

where t_{ij} denotes the disease type and x_{ij} the feature vector of the jth case in the ith clinic. Moreover we have available also a calibrative set of data C_{12}, assumed independent of D_1 and D_2 for reasons similar to those set out by Aitchison (1977a), with

$$C_{12} = \{(y_{1j}, y_{2j}); j = 1, \ldots, n\},$$

where y_{1j} and y_{2j} are associated measurements on the jth calibrative case by the methods of measurement in clinics 1 and 2 respectively.

As explained by Aitchison (1977a), model-building in this particular area requires careful consideration of the circumstances of the collection methods of the diagnostic training sets and the calibrative data. Rather than turn this paper into a philosophical discussion resulting in a long catalogue of models to cover the complete range of possibilities we concentrate on one particular structure which fits the application we have in mind. This involves the use of the diagnostic paradigm (Dawid, 1976), where effort is concentrated on the conditional distribution of type for given feature vector, in contrast with the sampling paradigm approach of Aitchison (1977a), where emphasis is laid on the assumed stability of the conditional distribution of the feature vector for given type. Since, as we shall see later, the clinic amalgamation model contains the system transfer model as a special case this paper thus provides the opportunity of describing a diagnostic paradigm alternative to the sampling paradigm version of system transfer of Aitchison (1977a).

Our objective is, in general, to provide each clinic with a diagnostic system for use with its own method of measurement. Since modelling of the links between the two clinics depends on the nature of the calibration experiment we take this aspect as our starting point. If the calibration experiment is a natural one in terms of the definition of Aitchison & Dunsmore (1975, p. 184) we have sufficient information to adopt a symmetrical approach, postulating conditional parametric models $p(x_1|x_2, \gamma_1)$ and $p(x_2|x_1, \gamma_2)$ for calibrating from x_2 to x_1 and from x_1 to x_2, respectively, where $\gamma_1 \in \Gamma_1$, $\gamma_2 \in \Gamma_2$ are the indexing parameters for the two classes of calibrative models. Because of this symmetry we need only show the construction of a diagnostic system for clinic 1. Clinic 1 wishes to relate disease type t to its own feature measurement x_1 through a diagnostic paradigm $p(t|x_1, \delta_1)$, where $\delta_1 \in \Delta_1$ is the indexing parameter of the class of diagnostic models. For example, a typical parametric model often advocated is the logistic discriminant model which sets

$$\text{pr}(t = 1|x_1, \delta_1) = 1 - \text{pr}(t = 2|x_1, \delta_1) = \exp(\delta_1' x_1)/\{1 + \exp(\delta_1' x_1)\}. \tag{2·1}$$

To use the diagnostic data D_2 from clinic 2 for the construction of the diagnostic system for clinic 1 we require to obtain, from the calibrative and diagnostic models for clinic 1, namely $p(x_1|x_2, \gamma_1)$ and $p(t|x_1, \delta_1)$, an induced model

$$p(t|x_2, \gamma_1, \delta_1) = \int_{X_1} p(t|x_1, \delta_1) p(x_1|x_2, \gamma_1) dx_1 \tag{2·2}$$

for the explanation of the variability of the data D_2 in terms of the clinic 1 parameters γ_1 and δ_1.

We can then focus our attention on the likelihood function $L(\gamma_1, \delta_1 | C_{12}, D_1, D_2)$ for γ_1 and δ_1 for the given calibrative and diagnostic data C_{12}, D_1, D_2:

$$L(\gamma_1, \delta_1 | C_{12}, D_1, D_2) = \prod_{j=1}^{n_1} p(t_{1j} | x_{1j}, \delta_1) \prod_{j=1}^{n_2} p(t_{2j} | x_{2j}, \gamma_1, \delta_1) \prod_{j=1}^{n} p(y_{1j} | y_{2j}, \gamma_1)$$

$$= L_1(\delta_1) L_2(\gamma_1, \delta_1) L_3(\gamma_1) \tag{2·3}$$

in an abbreviated notation which emphasizes the extent of the dependence of the three factors on the parameter components γ_1 and δ_1.

Adopting a Bayesian predictive approach for the reasons set out by Aitchison, Habbema & Kay (1977) we may resolve the statistical problem of making assessments of the diagnostic probabilities within clinic 1 as follows. First from the likelihood and with, if necessary, an almost vague prior on γ_1 and δ_1, we obtain the posterior distribution $p(\gamma_1, \delta_1 | C_{12}, D_1, D_2)$ for γ_1 and δ_1. Then, for a new case of unknown type but with known feature vector x_1 measured in clinic 1, we compute the diagnostic assessment

$$p(t | x_1, C_{12}, D_1, D_2) = \int_{\Delta_1} p(t | x_1, \delta_1) p(\delta_1 | C_{12}, D_1, D_2) \, d\delta_1, \tag{2·4}$$

where the marginal density function $p(\delta_1 | C_{12}, D_1, D_2)$ is obtained by integrating out γ_1 in the full posterior distribution. The provision of an appropriate system for clinic 2 follows exactly the same procedure with x_1 and x_2 interchanged and γ_2 and δ_2 replacing γ_1 and δ_1.

If the calibration experiment is not natural but designed, so that we can consider only one of the conditional models, say $p(x_1 | x_2, \gamma_1)$, then the above symmetrical treatment is impossible. The diagnostic system for clinic 1 remains (2·4), but we have to be content with a much more indirect method of arriving at a diagnostic system for clinic 2. The induced diagnostic model (2·2), previously used only to make clinic 2 data available to clinic 1, has now, because of the absence of information on the conditional model $p(x_2 | x_1, \gamma_2)$, to serve as the basis of arriving at a diagnostic assessment

$$p(t | x_2, C_{12}, D_1, D_2) = \int_{\Gamma_1} \int_{\Delta_1} p(t | x_2, \gamma_1, \delta_1) p(\gamma_1, \delta_1 | C_{12}, D_1, D_2) \, d\gamma_1 \, d\delta_1 \tag{2·5}$$

for a new case of unknown type but with known feature vector x_2 measured in clinic 2.

Note that when we do have a natural calibration experiment we would not use (2·5) as an alternative to the clinic 2 counterpart of (2·4), since (2·5) does not then use all the available calibrative information. For this reason also we would not expect the two assessments to coincide. In the interests both of good statistical practice and of avoiding awkward doubly-multiple integrals of the form (2·5) we should choose natural calibration experiments whenever possible.

3. The normal linear calibrative-diagnostic model

We now consider the implications of adopting particular parametric forms for the calibration and diagnostic components of our model. For the diagnostic paradigm for clinic 1 we adopt the normal distribution function form with argument a linear form of the feature vector:

$$\text{pr}(t = 1 | x_1, \delta) = 1 - \text{pr}(t = 2 | x_1, \delta) = \Phi(\delta' x_1), \tag{3·1}$$

where Φ is the standard univariate normal distribution function and allowing the first component of x_1 to be 1 for the usual purpose of simplified notation and yet recognizing the

necessity of a constant term in the linear form $\delta' x_1$. Note that for simplicity we have now dropped the suffix notation in the parameters γ and δ. This model, identical in form to an additive multistimulus version of probit analysis, is preferred to the more popular logistic model (2·1) because its convolution-integral properties give the advantage of simple evaluation of (2·2). A detailed comparison of such logistic and normal diagnostic paradigms in a different context is given by Lauder (1978).

For the calibrative paradigm we adopt the normal linear regression model

$$p(x_1 | x_2, \gamma) = \phi(x_1 | Ax_2, B), \tag{3·2}$$

where $\phi(. | \mu, \Sigma)$ is an appropriately dimensioned multivariate normal density function with mean μ and covariance matrix Σ, and again allowance is made for a regression constant in the presence of a unit in the first component of x_2. With these particular normal linear forms (3·1) and (3·2) the awkwardness of modelling, namely the multiple integration (2·2) involved in formulating the induced diagnostic paradigm for clinic 2, is easily resolved by reduction to a simple one-dimensional integral through the transformation $v = \delta' x_1$, giving

$$\text{pr}(t = 1 | x_2, \gamma, \delta) = \int_{-\infty}^{\infty} \Phi(v)\, \phi(v | \delta' A x_2, \delta' B \delta)\, dv = \Phi(\varepsilon' x_2), \tag{3·3}$$

where $\varepsilon = A'\delta / \sqrt{(1 + \delta' B \delta)}$. Thus the induced diagnostic paradigm for clinic 2 is also of the normal linear form (3·1) with parameter ε instead of δ.

The simple forms of the factors (3·1)–(3·3) provide an easily computable likelihood function from which, by the Newton–Raphson iterative technique, expressible in a modified probit analysis form, we can arrive at maximum likelihood estimates (c, d) for (γ, δ) and also at the information matrix inverse J. The details are omitted here. By the standard Bayesian counterpart of maximum likelihood large-sample theory and in order to take advantage of the predictive over the estimative approach we adopt the approximate posterior forms $p(\gamma, \delta | C_{12}, D_1, D_2) = \phi(\gamma, \delta | c, d; J)$ and $p(\delta | C_{12}, D_1, D_2) = \phi(\delta | d, G)$, where G is the appropriate submatrix of J.

For a new case of unknown type but with known feature vector x_1 measured in clinic 1 we have, by (2·4), the diagnostic assessment

$$\text{pr}(t = 1 | x_1, C_{12}, D_1, D_2) = \int_{\Delta} \Phi(\delta' x_1)\, \phi(\delta | d, G)\, d\delta = \Phi\{d' x_1 / \sqrt{(1 + x_1' G x_1)}\}. \tag{3·4}$$

When the calibration experiment is a natural one diagnostic assessments for new cases in clinic 2 take a form similar to (3·4) by the symmetrical analysis of § 2. When the calibration experiment allows only calibration from x_2 to x_1 then for new cases in clinic 2 we would have to resort to the diagnostic assessment (2·5) which for the normal linear model takes the form

$$\text{pr}(t = 1 | x_2, C_{12}, D_1, D_2) = \int_{\Gamma} \int_{\Delta} \Phi\left\{ \frac{\delta' A' x_2}{\sqrt{(1 + \delta' B \delta)}} \right\} \phi(\gamma, \delta | c, d; J)\, d\gamma\, d\delta, \tag{3·5}$$

where $\gamma = (A, B)$. This multiple integral is of complex structure and requires for its evaluation numerical or Monte Carlo methods, reinforcing the soundness of the advice of § 2 to perform a natural calibration experiment wherever possible.

4. AN ILLUSTRATIVE COMPARISON WITH SIMPLIFIED METHODS

We can easily illustrate the appreciable differences in diagnostic probability assessments that may arise between the use of the full clinic amalgamation model as described above and

other simplified methods which ignore certain important aspects of the situation in terms of the illustrative data set of Table 1 involving a univariate feature. For the diagnosis of new cases in clinic 1 three main alternative methods are of particular interest for comparison purposes.

Table 1. *Diagnostic and calibrative data for the illustrative example*

Clinic 1, data D_1		Clinic 2, data D_2		Calibration, data C_{12}	
t_{1j}	x_{1j}	t_{2j}	x_{2j}	y_{1j}	y_{2j}
1	2·04	1	1·69	0·45	0·72
1	−0·13	1	0·60	0·40	1·40
2	0·22	1	0·18	−0·36	−1·06
2	−0·49	2	0·38	−0·65	−0·31
2	−0·85	2	−0·14		
2	−0·92	2	−0·92		

(i) *The single relevant clinic method.* Construct a normal linear diagnostic system based on the diagnostic training set D_1 alone, thus avoiding any need to calibrate. This method may not be possible when the number of cases in clinic 1 is too small in relation to the dimension of the feature vector. For this particular set of data the method is applicable.

(ii) *The system transfer method.* Construct a normal linear diagnostic system based on the diagnostic training set D_2 alone and use the system transfer technique (Aitchison, 1977a) based on the calibrative data C_{12}, to allow application to the new cases in clinic 1. This method ignores the information in the diagnostic training set D_1 and so can be investigated as a special case of the clinic amalgamation model with D_1 the empty set.

(iii) *The naive calibration method.* Convert the features of the training set D_2 to corresponding point calibrates in clinic 1 to produce through this naive calibration (Aitchison, 1977a) an augmentation, say $\hat{D}_2(C_{12})$, to the clinic 1 training set D_1. From the combined training set $\{D_1, \hat{D}_2(C_{12})\}$ construct a clinic 1 diagnostic system based on the normal linear model and apply this directly to the new cases in clinic 1.

On theoretical grounds (i), (ii) and (iii) are all subject to criticism. Systems (i) and (ii) ignore completely the diagnostic information contained in one of the clinics and so offend the principle of using all available information. System (iii) in its use of naive calibration is open to all the criticisms of possibly bad misrepresentation made by Aitchison (1977a), in particular the effects arising from ignoring imprecisions in this form of calibration.

For a univariate feature it is simplest to show the differences between the methods diagrammatically since we can present the graphs of the probability of type 1 assigned by each of the four methods against the measured feature value in clinic 1. The anticipated consequences are clearly demonstrated. The system transfer method, making use only of the data D_2 and C_{12}, arrives at probability assessments which are naturally less firm, that is nearer the value $\frac{1}{2}$, than the full clinic amalgamation method which uses all of D_1, D_2 and C_{12}. On the other hand, the naive calibration method gives the appearance of providing firmer assessments than the full clinic amalgamation method. This firmness is, however, unjustified in the sense that the method is based on the false supposition that the naive calibrates are known precisely and without error. See Aitchison (1977a) for a more detailed criticism of this mistaken assumption. The merits of the single relevant clinic method are more difficult to assess. If the calibration process is not very precise as in this example then the additional information provided by D_2 through C_{12} towards the diagnostic process in clinic 1 may not provide firmer assessments for all cases than the use of D_1 alone. This is

borne out in Fig. 1 but we reemphasize that the clinic amalgamation model is principally directed towards situations where the data are insufficient to allow the application of the single relevant clinic method.

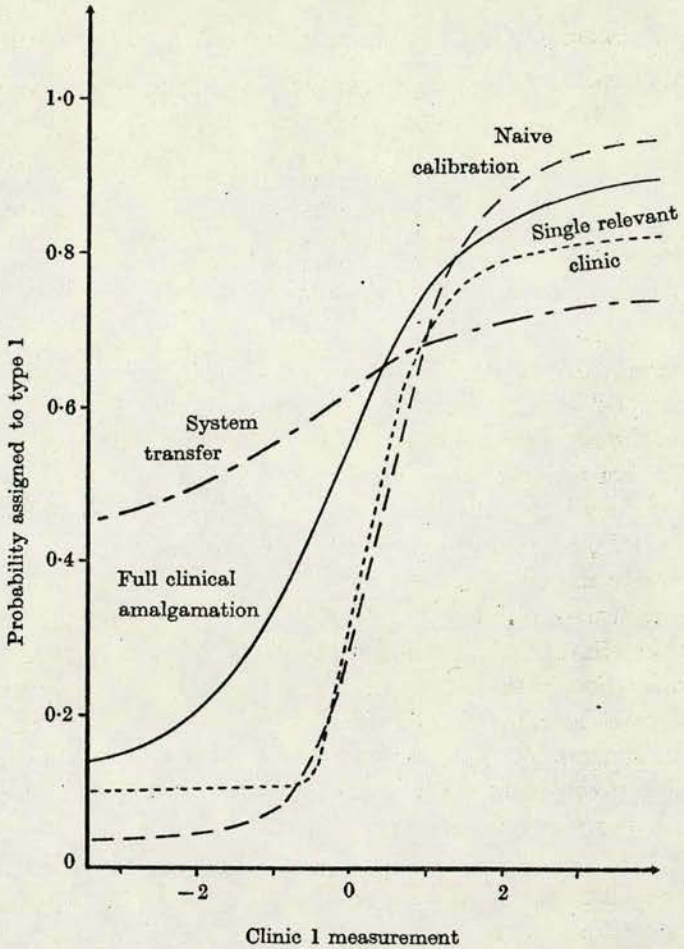

Fig. 1. Comparison of type 1 probabilities assigned by four different diagnostic systems.

A fourth, and more acceptable, alternative is to build up a diagnostic system for clinic 1 by an augmentation of the diagnostic training set D_2 but replacing each naive calibrate by a full calibrative density function, say $p(x_1|x_2, C_{12})$, as defined by Aitchison & Dunsmore (1975, p. 186). The augmented set then consists of two qualities of data, the original (t_{1j}, x_{1j}) with feature vectors x_{1j} $(j = 1, \ldots, n_1)$ known and the transferred cases $\{t_{2j}, p(x_{1j}|x_{2j}, C_{12})\}$ with the feature vectors not known exactly but with knowledge of the nature of their imprecision. Such training sets do not form the basis of any standard diagnostic analysis but a general method of statistical diagnosis based on features observed with variable precision has been developed recently in an unpublished report by J. Aitchison and I. J. Lauder, who in their illustrative examples provide an application in this area of calibrative diagnosis. Since the calibrative density function $p(x_1|x_2, C_{12})$ would be formed by the process

$$p(x_1|x_2, C_{12}) = \int_\Gamma p(x_1|x_2, \gamma)\, p(\gamma|C_{12})\, d\gamma,$$

there is some loss of information in separating the calibration problem from the diagnostic one, but the distortion is likely to be small compared with either the complete ignoring of D_2 or the naive calibration approach.

5. An application

Aitchison (1977a) showed how a statistical diagnostic system had to be adapted for application to cases in another clinic or in the same clinic where the methods of measurement of some of the features had been changed. The system transfer application there was to the area of the preoperative differential diagnosis of two types of Conn's syndrome and the transfer problem arose because the method of measurement of one of the features had been changed. Since the new method of measurement is now used on all new cases the designation of 'clinics' using the new and old methods of measurements as clinic 1 and 2, respectively, means that we need only aim for diagnostic assessments of the form (3·4) for new cases in clinic 1. The original training set D_2 in clinic 2 consists of 20 cases of type 1 and 11 cases of type 2 with an eight-dimensional feature vector, and clinic 1 has now, on the basis of cases diagnosed by the system transfer method of Aitchison (1977a) and with type subsequently confirmed histopathologically, another training set D_1 consisting of 17 cases of type 1 and 4 cases of type 2. For the illustrative purposes of this paper we used only the three most discriminating of the eight features, namely the plasma concentrations of potassium, renin and aldosterone, the last of which is the feature involved in the calibration aspect of the problem. This reduction of the feature vector also helped to avoid the awkwardness of the diagnostic paradigm which arises when there is complete hyperplane separation of the two types in the sense of Anderson (1972).

The calibration data C_{12} arising from a natural calibration experiment consist of 72 blood samples, each divided into two, one being determined by the old double-isotope method of clinic 2 and the other by the new radioimmunoassay method of clinic 1 with respect to one of the features, the concentration of aldosterone. We have thus (Aitchison, 1977a) a partial calibration problem and following the notation there we use superscript 1 to denote that part of the feature vector not requiring calibration and superscript 2 for that part requiring calibration. We then simply rewrite (3·1) as

$$\mathrm{pr}\,(t = 1 \,|\, x_1, \delta) = \Phi(\delta_0 + \delta_1' x_1^{(1)} + \delta_2 x_1^{(2)}) \tag{5·1}$$

and confine the calibration regression model to the appropriate component by writing

$$\mathrm{p}(x_1^{(2)} \,|\, x_2^{(2)}, \gamma) = \phi(x_1^{(2)} \,|\, \alpha + \beta x_2^{(2)}, \sigma^2). \tag{5·2}$$

Then (3·3) provides the basis for handling the diagnostic data D_2 in clinic 2 by becoming

$$\mathrm{pr}\,(t = 1 \,|\, x_2, \gamma, \delta) = \Phi[\{\delta_0 + \delta_1' x_2^{(1)} + \delta_2(\alpha + \beta x_2^{(2)})\}/\sqrt{(1 + \delta_2^2 \sigma^2)}]. \tag{5·3}$$

Because of the small number, only 4, of cases of type 2 in clinic 1, it was not considered sensible to attempt to apply the single relevant clinic method of § 4 in this situation. The full clinic amalgamation method, the system transfer method and the naive calibration method were, however, each applied to obtain diagnostic assessments for the 21 training cases in clinic 1, by resubstitution, and for 22 new cases of unknown type. Figure 2 provides, for each of these 43 cases, a comparison of the probabilities of type 1 assigned by the clinic amalgamation and the system transfer methods. Note that on the whole the full clinic amalgamation method gives assessments of greater firmness than the system transfer method, and in a number of cases the differences are substantial.

It is not possible to show the results of the naive calibration method on the same diagram since there is practically no difference from those of the full clinic amalgamation method, in contrast to our illustrative example of § 4, and also in contrast to the warnings against naive calibration by Aitchison (1977a). Some explanation of this contrast is called for. The explanation has two components. First, naive calibration here is directed towards producing a diagnostic system for clinic 1 in contrast to the assessment for a new case in clinic 1 as in

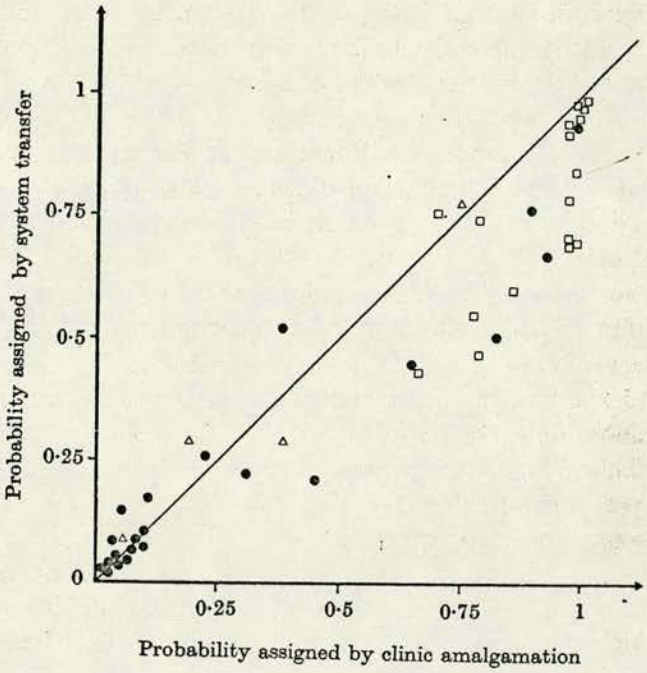

Fig. 2. Comparison of type 1 probabilities assigned by full clinic amalgamation and system transfer methods: □, clinic 1 case of known type 1; △, clinic 1 case of known type 2; ●, new case in clinic 1 of unknown type.

system transfer. The calibration experiment is sizeable and so produces reasonably reliable estimates of the regression parameters. Although for transfer of a single case the naive calibrate has an appreciable unreliability, ignoring this unreliability may yet produce, because of averaging over a number of cases, a satisfactory diagnostic system for clinic 1. Secondly, since new cases have features measured in clinic 1 there is no need for any additional calibration technique to be applied to the diagnostic assessment stage, in contrast to the situation in system transfer where each new case requires individual calibration.

Despite the fact that we seem to have succeeded with naive calibration in this particular application, we hope that the illustrative example of § 4 will serve as a reminder that there can be substantial differences. Unfortunately it is virtually impossible to give any simple set of instructions as to when we can resort to naive calibration. Such factors as the 'quality' of the calibration as discussed by Aitchison (1977a), the sizes of the diagnostic training sets, the extent of the diagnosibility of the types for the given diagnostic features, play a complicated and interrelated role as determinants. Since computations for the full clinic amalgamation method are not essentially any more difficult than those of naive calibration we would advocate playing safe and applying the full clinic amalgamation system.

6. Extensions and discussion

The theory and applications of the preceding sections have been confined to the case of two clinics and two types. Extension from two to k clinics with diagnostic training sets $D_1, \ldots, D_k$ can be reasonably straightforward provided that the calibration experiment is suitably designed. We briefly consider two versions.

(1) Suppose that the calibration experiment consists of n cases, independent of $D_1, \ldots, D_k$, measured in every clinic with resulting measurements of $x_1, \ldots, x_k$ for a typical case. Suppose further that the joint variability of $(x_1, \ldots, x_k)$ is modelled by the parametric class $p(x_1, \ldots, x_k | \gamma)$. Then all pairwise conditional distributions are available so that the symmetric analysis of § 2 can be applied. For example, starting with a diagnostic paradigm $p(t | x_1, \delta)$ for clinic 1 we obtain $k-1$ induced diagnostic models corresponding to (2·2), with

$$p(t | x_c, \gamma, \delta) = \int_{X_1} p(t | x_1, \gamma) p(x_1 | x_c, \gamma) \, dx_1$$

for clinic $c = 2, \ldots, k$. The likelihood for diagnosis in clinic 1 is then easily formed and, for normal linear diagnostic calibrative paradigms, gives normal linear diagnostic paradigms for each of the clinics. Though the computations are more complex because of the more complicated nature of γ, they are certainly feasible.

(2) Independent calibration experiments of each clinic with a central clinic, say clinic 1, could be modelled unsymmetrically by separate conditional models $p(x_1 | x_c, \gamma_c)$ relating clinic 1 to clinic c ($c = 2, \ldots, k$), with this conditional density formation replacing $p(x_1 | x_c, \gamma)$.

Extension from two to three types is straightforward. Where the univariate cumulative normal distribution function provided a mechanism for two types the corresponding standardized bivariate distribution function, say Φ_2, provides the mechanism for three types through the modelling for clinic 1:

$$\text{pr}(t = 1 | x_1, \delta) = \Phi_2(\delta_1' x_1, \delta_2' x_1),$$

$$\text{pr}(t = 2 | x_1, \delta) = \Phi_1(\delta_1' x_1) - \Phi_2(\delta_1' x_1, \delta_2' x_1),$$

$$\text{pr}(t = 3 | x_1, \delta) = 1 - \Phi_1(\delta_1' x_1).$$

See the unpublished report by J. Aitchison and I. J. Lauder for further details of the usefulness of this model. Along with the calibrative models already used it leads through (3·3) to factors for the likelihood involving at worst bivariate normal integrals over rectangular regions, whose computation is available through standard algorithms.

Extension to r types is conceptually straightforward but depends on the availability of methods for computing $(r-1)$-dimensional multivariate normal distributions over cuboidal regions of $(r-1)$-dimensional space. The logistic discriminant analysis alternative provides no means of escape since even for only two types the counterpart of integral (3·3) requires the use of approximate methods.

Finally, all the considerations of this paper could have been formulated in terms of the sampling paradigm, which concentrates attention on the conditional distribution of feature vector for given type, rather than the diagnostic paradigm we have used. Indeed the sampling paradigm has considerable advantages of tractability for more than three types. Particularly in situations where different clinics are involved, however, the use of the sampling paradigm is open to the substantial criticisms set out by Dawid (1976) and so we have preferred to place emphasis here on the diagnostic paradigm.

REFERENCES

AITCHISON, J. (1977a). A calibration problem in statistical diagnosis: The system transfer problem. *Biometrika* **64**, 461–72.

AITCHISON, J. (1977b). Calibration problems in statistical diagnosis. *Bull. Inst. Int. Statist.* **47**, 4, 9–12.

AITCHISON, J. & DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

AITCHISON, J., HABBEMA, J. D. F. & KAY, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statist.* **26**, 15–25.

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

DAWID, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647–58.

LAUDER, I. J. (1978). Computational problems in predictive diagnosis. In *Compstat* 1978, pp. 186–92. Vienna: Physica-Verlag.

[*Received September* 1978. *Revised November* 1978]

AITCHISON, J. and LAUDER, I.J. (1979)

Statistical diagnosis from imprecise data

Reprinted from *Biometrika* 66, 475-83

Biometrika (1979), 66, 3, *pp.* 475–83
Printed in Great Britain

Statistical diagnosis from imprecise data

BY J. AITCHISON AND I. J. LAUDER
Department of Statistics, University of Hong Kong

SUMMARY

The fact that diagnostic measurements are often subject to error, with the extent of the imprecision varying from case to case, is largely ignored in current methodology of statistical diagnosis. Models taking full account of such imprecision are proposed and the necessary methods developed. In particular, a useful combination of a cumulative-normal diagnostic model with a normal error model is studied. Applications to two specific medical diagnostic problems illustrate the differing extents of the misrepresentation that may be involved in the use of techniques that ignore imprecision.

Some key words: Calibration; Cumulative normal-normal model; Diagnostic paradigm; Logistic-normal model; Measurement error; Medical diagnosis; Sampling paradigm.

1. INTRODUCTION

The recorded features or measurements of cases arising in statistical diagnostic problems are often imprecise for a variety of reasons, such as physiological variability (Ferriss *et al.*, 1970), assessment by assay techniques (Cost & Vegter, 1962), calibration between sources such as different clinics (Aitchison, 1977) and subjective measurement or observer error (Buckton *et al.*, 1976; Smyllie, Blendis & Armitage, 1965).

Such imprecision in feature vectors, although often recognized, is seldom taken directly into account in the actual construction of a statistical diagnostic system. To the extent that calibration problems in diagnosis are a subclass of the problems studied here the considerable effects that neglect of known imprecision can have on clinical practice have already been shown by Aitchison (1977). For the simple univariate case with known normal distributions of features for given disease type, Good & Card (1971) point out that the effect of neglecting error can be appreciable.

The reasons for the neglect of imprecision in general are presumably

(i) the lack of appropriate statistical methods in this kind of discriminant analysis, and

(ii) the assumption that disregard of such imprecision has negligible consequences in practice.

Our purpose in this paper is first to remedy (i). It is relatively simple to model diagnostic problems for degrees of imprecision varying from component to component of the feature vector and from case to case. Only by so modelling is it possible to investigate the validity of the assumption (ii).

2. DIAGNOSTIC MODELS FOR IMPRECISE DATA

For diagnostic modelling purposes we follow Dawid (1976) and adopt the diagnostic paradigm, which concentrates on the conditional distribution of type for a given feature vector, in preference to the sampling paradigm, in which the conditional distribution of the feature vector for a given type plays the central role. For a case (t, x) of diagnostic type t belonging to a set T and with true feature vector x belonging to a set X, a parametric model

of the diagnostic paradigm specifies the conditional distribution $p(t|x,\delta)$, where δ is a vector parameter belonging to a set Δ. Instead of the true feature vector x we can observe only a possibly inaccurate vector y. We suppose that we can assess the conditional distribution $p(x|y)$ of x for the observed y. The parametric model with respect to the observable feature vector y then takes the form

$$p(t|y,\delta) = \int_X p(t|x,\delta)\,p(x|y)\,dx. \tag{2·1}$$

Information about the unknown parameter δ is obtained from a training set

$$D = \{(t_i, y_i);\ i = 1,\ldots,n\}$$

consisting of the known type and observed feature vector for each of n cases. We suppose that this information can be summarized in terms of a posterior distribution $p(\delta|D)$ so that the practical advantages and greater realism of the predictive diagnostic approach (Aitchison, 1976; Aitchison, Habbema & Kay, 1977) can be exploited. For a new case of unknown type t and observed feature vector y the predictive diagnostic probabilities are

$$p(t|y,D) = \int_\Delta p(t|y,\delta)\,p(\delta|D)\,d\delta \tag{2·2}$$

$$= \int_\Delta \int_X p(t|x,\delta)\,p(x|y)\,p(\delta|D)\,dx\,d\delta. \tag{2·3}$$

Three components of the statistical problem can be recognized.

(1) The likelihood problem: for data D the likelihood of δ is

$$L(\delta|D) = \prod_{i=1}^{n} p(t_i|y_i,\delta)$$

and so for an effective analysis we clearly require a ready means of evaluating integrals of the form (2·1).

(2) The posterior distribution problem: the complicated form of the likelihood rules out the possibility, whatever the nature of any prior distribution $p(\delta)$ over Δ we care to invent, of obtaining a neat, tractable form for $p(\delta|D)$. The problem is thus to decide what additional simplifying assumptions will provide a balance between realism and the development of an operational tool.

(3) The diagnostic problem for a new case: mathematically this involves obtaining numerical values for multiple integrals of the form (2·2) or 'doubly multiple' integrals of the form (2·3).

3. THE CUMULATIVE NORMAL-NORMAL AND THE LOGISTIC-NORMAL MODELS FOR TWO TYPES

For the remainder of this paper we confine attention to continuous features. In this section we restrict consideration to situations where diagnosis between only two types, 1 and 2, is required; possible extensions to more than two types are discussed in § 5.

As parametric model of the diagnostic paradigm we adopt the cumulative normal form

$$\mathrm{pr}(t=1|x,\delta) = 1-\mathrm{pr}(t=2|x,\delta) = \Phi(\delta'x), \tag{3·1}$$

where Φ is the standard univariate normal distribution function, in preference to the more

popular logistic discriminant form

$$\text{pr}(t=1|x,\delta) = 1-\text{pr}(t=2|x,\delta) = e^{\delta' x}/(1+e^{\delta' x}). \tag{3·2}$$

The great advantage of (3·1) over (3·2) is undoubtedly the availability of a parametric form of the error paradigm which allows the explicit evaluation of the convolution-type integrals (2·1). For if

$$p(x|y) = \phi(x|By, S), \tag{3·3}$$

where $\phi(.|\mu,\Sigma)$ is the density function of a multivariate normal distribution with mean μ and covariance matrix Σ, then (2·1) based on (3·1) and (3·3) can be expressed as

$$\Phi(x|\delta' By, 1+\delta' S\delta). \tag{3·4}$$

We know of no form $p(x|y)$ which in combination with (3·2) gives such an explicit evaluation of (2·1). Moreover, in terms of another important aspect, namely quality of fit, there is seldom any significant difference in terms of separate family comparisons (Cox, 1962) between the two forms. Mainly on grounds of tractability we therefore favour form (3·3). For a study of the computational problems associated with the normal and logistic forms, see Lauder (1978).

Form (3·3) with $B = I$ is appropriate if y is quoted as an estimate of x with computed variance S or, in the multivariate situation, estimated covariance matrix S. We have retained in (3·3) the more general B since information about x is sometimes obtained through some indirect form of measurement such as calibration and assay.

Note that from (3·4) the simply established inequality

$$|p(t|y,\delta)-\tfrac{1}{2}| < |p(t|x=By,\delta)-\tfrac{1}{2}|\ . \tag{3·5}$$

verifies the intuitive modelling requirement that knowledge of inaccurate y rather than true x must lead us to diagnostic probabilities which are closer to $\frac{1}{2}$, the diagnostic assessment expressing the greatest uncertainty.

In general, individual cases of the training set will have B_i and S_i $(i = 1,\ldots,n)$ differing from each other and from the B and S of a new case, so that the precise form of the diagnostic model (3·4) varies from case to case. The likelihood problem (1) of § 2 is, however, easily resolved for the cumulative-normal model since the integrals of form (2·1) take the explicit and easily computable forms (3·4). It is tempting to hope that the problems of taking account of imprecision are thereby automatically resolved but we shall see in § 4 that imprecision in clinical situations can cause substantial, and at times insurmountable, further difficulties.

To arrive at diagnostic assessments for new cases through the evaluation of integrals of the form (2·2) or (2·3) some simple form for $p(\delta|D)$ must be obtained. We assume the applicability of the Bayesian form of large-sample maximum likelihood theory (Lindley, 1961); in other words, δ is assumed to be approximately multivariate normally distributed as $N\{\hat{\delta}, V(\hat{\delta})\}$, where $\hat{\delta}$ is the maximum likelihood estimate and $V(\hat{\delta})$ the usual asymptotic estimate of the covariance matrix of $\hat{\delta}$, evaluated at $\hat{\delta}$. The algorithm to obtain $\hat{\delta}$ and $V(\hat{\delta})$ by the Newton–Raphson method is only slightly more complicated than a straightforward probit analysis, each iterative step being expressible in weighted regression form. Write

$$u_i = \delta' B_i y_i/\sqrt{(1+\delta' S_i\delta)}, \quad \omega = \phi^2/\{\Phi(1-\Phi)\}; \tag{3·6}$$

define the weights w_i as $\omega(u_i)$, and the regressor vector X_i and the regressand Y_i by

$$X_i = (1+\delta' S_i\delta)^{-3/2}\{(1+\delta' S_i\delta)\,B_i y_i - \delta' B_i y_i S_i\delta\}, \quad Y_i = u_i + \{2-t_i-\Phi(u_i)\}/\phi(u_i).$$

The iterative relation determining the rth iterate $\delta^{(r)}$ is then

$$\delta^{(r)} = (\Sigma_i w_i X_i X_i')^{-1}(\Sigma_i w_i X_i Y_i), \tag{3·7}$$

where the right-hand side is evaluated at the $(r-1)$th iterate $\delta^{(r-1)}$. At convergence $\hat{\delta} = \delta^{(r)}$ and $V(\hat{\delta})$ is the inverse matrix, the first factor of the right-hand side.

This resolution of the posterior distribution problem (2) allows us to proceed to problem (3) of § 2, the evaluation of the integral

$$\text{pr}\,(t = 1 \mid y, D) = \int_\Delta \Phi\{\delta' By/\sqrt{(1+\delta' S\delta)}\}\,\phi\{\delta \mid \hat{\delta}, V(\hat{\delta})\}\,d\delta. \tag{3·8}$$

For a new case with an exact feature vector, and so with $S = 0$, this multiple integral can be evaluated explicitly as

$$\Phi[\hat{\delta}' By/\sqrt{\{1 + y'B'V(\hat{\delta})\,By\}}], \tag{3·9}$$

but for a new case with $S \neq 0$ there is no closed form for (3·8). Since numerical integration of such multivariate integrals is difficult and since it is possible to simulate an independent sequence $\{\delta_n\}$ of $N\{\hat{\delta}, V(\hat{\delta})\}$ vectors, Monte Carlo techniques are appropriate.

The Monte Carlo technique of importance sampling (Hammersley & Handscomb, 1964, p. 57) adapted for the evaluation of (3·8) by Lauder (1978) can be improved by the control variate technique (Hammersley & Handscomb, 1964, p. 59) in the following way. We can reexpress (3·8) as the sum of an explicit, easily computed, control term

$$\Phi[\hat{\delta}' By/\sqrt{\{1 + \hat{\delta}' S\hat{\delta} + y'B'V(\hat{\delta})\,By\}}] \tag{3·10}$$

and the integral

$$\int F(\delta)\,\phi\{\delta \mid \hat{\delta}, V(\hat{\delta})\}\,d\delta, \tag{3·11}$$

where

$$F(\delta) = \Phi\{\delta' By/\sqrt{(1+\delta' S\delta)}\} - \Phi\{\delta' By/\sqrt{(1+\hat{\delta}' S\hat{\delta})}\},$$

the integral (3·11) being approximated by importance sampling. In our experience of cases so far the contribution from the integral (3·11) has been negligible to two significant digits, so that the control term (3·10) itself may prove to be an adequate approximation in practice. If there is a need to improve upon this approximation one possibility is to note that, in a series expansion of $F(\delta)$ about $\hat{\delta}$, the first nonzero term arises from the Hessian matrix, $H(\delta)$ say, of second order derivatives of F, and can be expressed as $\frac{1}{2}\,\text{tr}\,\{H(\hat{\delta})\,V(\hat{\delta})\}$.

4. APPLICATIONS

4·1. *Preliminaries*

The two objectives of this paper are to draw attention to the possible consequences of ignoring imprecision in features and to illustrate the simple methodology of § 3 which takes full account of such imprecision. One of our motivating problems, which is the diagnosis, and differential diagnosis of four forms, of Cushing's syndrome, is too complex to present in detail here and so we have selected the simplest subproblem which will fulfil these objectives.

4·2. *Differential diagnosis of Cushing's syndrome*

Cushing's syndrome is a condition involving high blood pressure with involvement of the adrenal glands, and we confine attention here to the problem of differentiating between its two benign forms, adrenal adenoma, type 1, and adrenal hyperplasia, type 2. Such differentiation is of practical importance because the treatments are quite different for the two types.

A training set of 7 type 1 and 27 type 2 cases is available, each case, for our limited illustration, having a two dimensional feature vector, consisting of urinary excretion rates of two steroid metabolites, allo-tetrahydrocortisol and tetrahydrocortisone, determined by the paper chromatography method of Cost & Vegter (1962), who state that their method has a 20% coefficient of variation. To take account of a coefficient of variation equal to c the error model (3·3) takes the form

$$p(x|y,S) = \phi[x|y,\{c\,\mathrm{diag}\,(y)\}^2], \tag{4·1}$$

where $\mathrm{diag}\,(y)$ is the diagonal matrix whose diagonal elements are the components of the vector y.

We reemphasize that our main purpose here is to investigate for new cases the extent to which admission of this factor of imprecision alters the diagnostic assessments which we would obtain using the excretion rates as if they were precise. A general study of the effects of imprecision requires the evaluation of the multiple integral (3·8) for various degrees of imprecision in the training set and in the new case. To avoid any possible confounding of imprecision effects with the accuracy of the approximation (3·10), we can conveniently approach the general study in two stages. At the first stage we ask what is the effect of recognizing imprecision only in the training set, with $S = 0$ for a new case, when (3·8) takes the exact form (3·9). The second stage is then simply to let S increase from zero and to use (3·10) or a Monte Carlo technique.

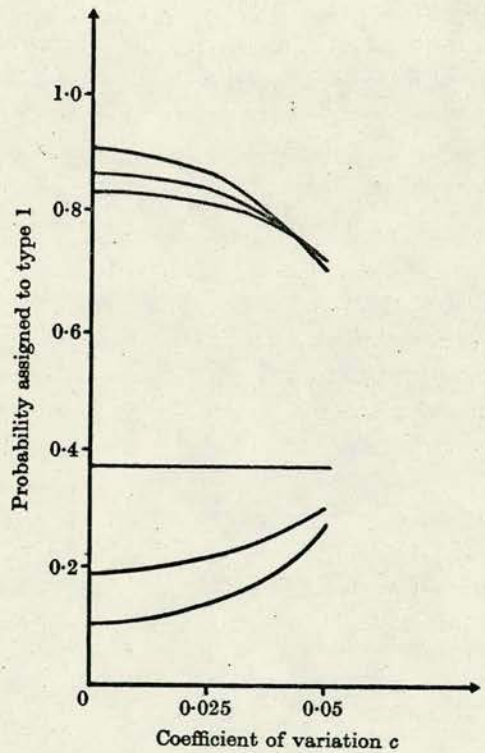

Fig. 1. Changes in the diagnostic probabilities of typical cases as the coefficient of variation increases.

To study the first stage, model (3·4) with error distribution (4·1) was fitted by the iterative procedure (3·7) with coefficient of variation $c = 0$, that is considering the training set as accurate, and then with increasing values of c up to 5%. For each c, after convergence, the

predictive diagnostic probabilities (3·9) were determined for 40 new cases. For all cases but one the diagnostic probabilities changed substantially as illustrated by typical cases in Fig. 1, the probabilities moving towards $\frac{1}{2}$.

At the second stage we investigated the additional effect of increasing the coefficient of variation for a new case up to 5% using the approximation (3·10). The further reduction in the diagnostic probabilities was smaller than 2% in all cases.

Our investigation has been restricted to a maximum coefficient of variation of 5% instead of the actual 20%. As the coefficient of variation increases past 5% it becomes increasingly difficult to obtain convergence by the Newton–Raphson method. We return to this difficulty in § 5.

4·3. *A two-clinic problem in differential diagnosis*

A second application involves an alternative to the clinic amalgamation problem discussed by Aitchison (1979). Here, two clinics wish to amalgamate their training sets but one of the features, plasma concentration of aldosterone, is determined by different techniques in the two clinics. In presenting this example to illustrate the new methods, we have used the same three features for the differential diagnosis of the two types of Conn's syndrome as used by Aitchison (1979). In brief, the training set is here considered to consist of 21 cases from clinic 1, with 17 of type 1 and 4 of type 2, whose feature measurements, by the latest method, are regarded as exact; and 31 cases from clinic 2 with 20 of type 1 and 11 of type 2, whose first feature, because of its calibration to the first clinic standard, is imprecise but whose other two feature measurements are exact. The B_i and S_i of § 3 then take the forms

$$B_i = I_4, \quad S_i = 0 \quad (i = 1, \ldots, 21),$$

$$B_i = \begin{bmatrix} a & b & 0 \\ 0 & 0 & I_2 \end{bmatrix}, \quad S_i = \operatorname{diag}[0, c\{D + E(y-F)^2\}, 0, 0] \quad (i = 22, \ldots, 52),$$

where a and b are the calibration regression coefficients, c is the residual mean square of the calibration regression and y is the aldosterone measurement in clinic 2; Aitchison & Dunsmore (1975, § 2·5) and Aitchison (1977, p. 470) give further details. Our approach here thus completely separates out the calibration problem from the diagnostic one, producing a single training set with a mixture of precise cases and imprecise, calibrated cases, whereas Aitchison (1979) retains the two training sets using a calibration paradigm as the binding element. Where the calibration experiment is large, as in the present problem, the two approaches are likely to give similar results.

Here new cases are measured by the new technique so that $B = I_4$, $S = 0$ and only the first stage defined in § 4·2, involving the effects of imprecision in the training set, need be considered. The degree of imprecision in this example is represented by c and so we can again study the effects of ignoring the imprecision by setting $c = 0$ as well as to its actual value $c = 0{\cdot}184$. For both assumptions about c, there were no practical problems of numerical convergence in fitting the cumulative normal model (3·4), and (3·9) was applied to obtain the predictive diagnostic probabilities for 22 new cases. Our findings are identical with those of Aitchison (1979). The differences between the odds assigned on a basis of ignoring imprecision and those taking account of the calibration imprecision are negligible. This is in sharp contrast to the substantial differences in the application to Cushing's syndrome. A possible explanation is that precise cases are sufficiently frequent in the combined training set to prevent the imprecise, calibrated cases from causing much of an effect at their actual degree of imprecision. If, however, we let the degree of imprecision increase well beyond its actual value, up to

$c = 1$, the odds do change by a factor of 3 in a substantial number of cases. From this result it is fairly safe to conclude that it is the frequency and the magnitude of imprecision that dually affect the assessments.

5. Discussion

Our illustrative examples have demonstrated that ignoring imprecision can give diagnostic assessments with a false appearance of firmness. Explicit recognition of imprecision can be incorporated in diagnostic modelling and has the expected effect of reducing the firmness of the diagnostic assessments. This effect can be so extreme that diagnostic assessments are practically the same whatever the feature vector. If the imprecision is appreciable standard procedures, such as Newton–Raphson iteration, which work readily under the assumption of precise data, fail completely for the degree of imprecision actually present.

This effect of imprecision can be easily demonstrated by a simple illustrative model. Suppose that there is a one dimensional feature, a one dimensional parameter δ with model $\mathrm{pr}(t = 1 \mid y, \delta) = \Phi(\delta y)$, and that the training set D consists of four cases

$$(2, -1), \quad (1, -0{\cdot}5), \quad (2, 0{\cdot}5), \quad (1, 2)$$

with $S_i = s^2$ $(i = 1, \ldots, 4)$. Figure 2 shows the graphs of the log likelihood for $s = 0$, 1, 1·4, 1·8, and illustrates two points of difficulty. First, a maximum likelihood estimate may not even

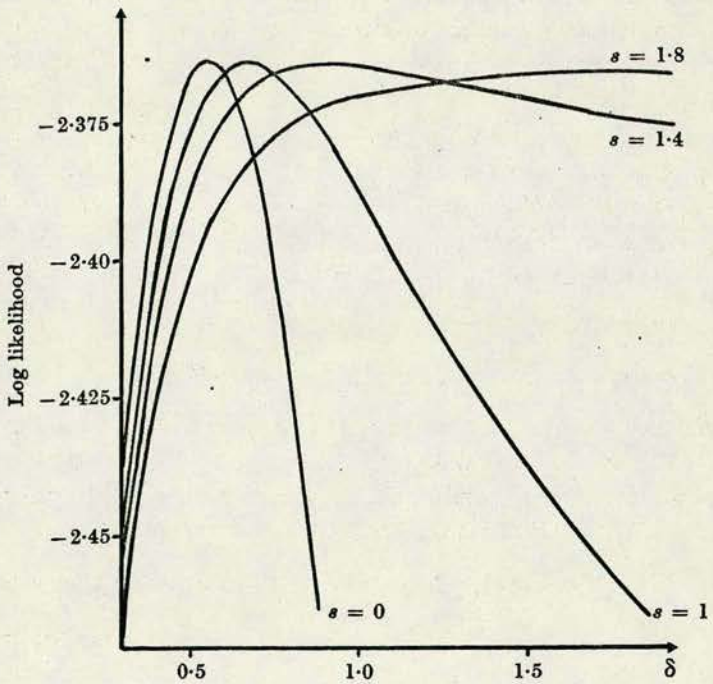

Fig. 2. Log likelihood graphs for illustrative example of § 5.

exist for some degrees of imprecision. Secondly, even when a maximum likelihood estimate exists it is clear that as s increases and the graphs flatten the determination of $\hat{\delta}$ becomes a more and more difficult numerical task and, moreover, the asymptotic normal approximation of maximum likelihood theory must eventually deteriorate. These difficulties obviously can persist for higher dimensional δ, as we have encountered in the application of § 4·1.

Two aspects of our modelling suggest that it may be interesting to consider the role of the EM algorithm of Dempster, Laird & Rubin (1977) for maximum likelihood estimation from incomplete data. First, the observable y is incomplete in the sense of being an imprecise form of the unobservable and precise x. Secondly, the similarity of the cumulative normal diagnostic model to probit analysis with its connotation of an unobservable tolerance which determines the type of binary response suggests that we might regard type t as incomplete in relation to a latent indicator variable v.

More specifically, define a complete pair (v, x) with $p(v|x,\delta)$ of $N(\delta'x, 1)$ form and related to incomplete (t, y) through a conditional distribution $p(x|y)$ of $N(By, S)$ form and the rule that $t = 1$ if $v > 0$ and $t = 2$ if $v \leqslant 0$. Then

$$\text{pr}\,(t = 1|y,\delta) = \int_X p(v > 0|x,\delta)\,p(x|y)\,dx = \Phi\{\delta'By/\sqrt{(1+\delta'S\delta)}\},$$

and so (v, x) bears the relationship of complete data for the diagnostic model (3·4) with observable data (t, y). The EM algorithm for its rth step then takes the following simple form.

E step: with given $\delta^{(r)}$ compute, taking + for $t = 1$ and − for $t = 2$,

$$v_i^{(r)} = \delta^{(r)\prime} y_i \pm \sqrt{(1+\delta^{(r)\prime} S_i \delta^{(r)})}\,h(\pm u_i^{(r)}), \quad x_i^{(r)} = y_i \pm (1+\delta^{(r)\prime} S_i \delta^{(r)})^{-\frac{1}{2}} S_i \delta^{(r)} h(\pm u_i^{(r)}),$$

where h is the function ϕ/Φ and u_i is as defined in (3·6).

M step: regress $v_i^{(r)}$ on $x_i^{(r)}$ to obtain $\delta^{(r+1)}$ as regression coefficients.

The algorithm, though attractively simple, is disappointing in its convergence properties (Lauder, 1978), not surprisingly in a situation where Newton–Raphson iteration can be ineffective. The concept of a latent indicator variable v essential to this EM algorithm approach does, however, open up ways of overcoming the difficulty of extending the modelling to more than two types. For k types we consider a $(k-1)$-dimensional indicator vector v for which $p(v|x)$ is $N_{k-1}(\Lambda x, I)$. To produce a specific diagnostic model all that is required is to partition the $(k-1)$-dimensional v-space into k regions R_t $(t = 1, \ldots, k)$, each associated with one of the types. For example, the standard version of logistic discrimination for precise feature vectors, as given by Anderson (1972), essentially uses

$$R_t = \begin{cases} \{v: v_t \geqslant 0,\ v_t = \max(v_i) \quad (i = 1, \ldots, k-1)\} & (t = 1, \ldots, k-1), \\ \{v: v_i < 0 \quad (i = 1, \ldots, k-1)\} & (t = k). \end{cases}$$

To a large extent the method of partitioning through the use of $k-1$ linear forms Λx is arbitrary and dictated by the tractability of the ensuing analysis. Such a view encourages the exploration of other forms of partition. The partition suggested by Aitchison (1979) for the extension of the clinic amalgamation problem to $k = 3$ uses

$$R_1 = \{v;\ v_1 < 0, v_2 < 0\}, \quad R_2 = \{v: v_1 < 0, v_2 \geqslant 0\}, \quad R_3 = \{v: v_1 \geqslant 0\}$$

but considers the components of v correlated so that bivariate normal distribution functions are involved. With a partition of this type the added assumption adopted above that, for given x, the v are uncorrelated yields, for any k, expressions for $p(t|x,\delta)$ involving only products of univariate Φ functions. This allows for tractable combination with an error paradigm and with computations no greater in complexity than for $k = 2$. We are currently exploring this very simple extension.

In this paper we have concentrated on the diagnostic paradigm. Modelling for the sampling paradigm is also straightforward if we take $p(x|t,\delta)$ as $N(\mu_t, \Sigma_t)$ and the error model, now conditionally $p(y|x)$, in the form $N(Bx, S)$. Then so far as the observable feature vector y is

concerned the sampling model takes the form

$$p(y|t,\delta) = \int_X p(y|x)p(x|t,\delta)\,dx = N(B\mu_t, S + B\Sigma_t B),$$

so that the likelihood can be obtained explicitly. The subsequent posterior distribution problem and the diagnostic problem for a new case are resolved in much the same manner as for the diagnostic paradigm except that to convert $p(x|t,D)$ to $p(t|x,D)$ through Bayes's formula we require to insert an assumption concerning the incidence rate for a new case.

There are clearly many unanswered questions raised by the explicit modelling of imprecision in statistical diagnosis. Two related questions which would seem worth early consideration are the following. To what extent should knowledge of imprecision affect some of the standard methods of feature selection? To what extent does the retention of features which would be discarded under a selection procedure which ignores imprecision features provide a necessary redundancy to compensate for imprecision effects?

References

Aitchison, J. (1976). Goodness of prediction fit. *Biometrika* **63**, 547–54.

Aitchison, J. (1977). A calibration problem in statistical diagnosis: The system transfer problem. *Biometrika* **64**, 461–72.

Aitchison, J. (1979). A calibration problem in statistical diagnosis: The clinic amalgamation problem. *Biometrika* **66**, 357–66.

Aitchison, J. & Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

Aitchison, J., Habbema, J. D. F. & Kay, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statist.* **26**, 15–25.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Buckton, K. E., O'Riordan, M. L., Jacobs, P. A., Robinson, J. A., Hill, R. & Evans, H. J. (1976). *C*- and *Q*-band polymorphism in the chromosomes of 3 human populations. *Ann. Hum. Genet.* **40**, 99–112.

Cost, W. S. & Vegter, J. J. M. (1962). Quantitative estimation of adrenocortical hormones and their α-ketolic metabolites in urine. *Acta Endocr.* **41**, 571–83.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B **24**, 406–24.

Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647–58.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Ferriss, J. B., Brown, J. J., Fraser, R., Kay, A. W., Neville, A. M., O'Muircheartaigh, I. G., Robertson, J. I. S., Symington, T, & Lever, A. F. (1970). Hypertension with aldosterone excess and low plasma-renin: Preoperative distinction between patients with and without adrenocortical tumour. *The Lancet* **2**, 995–1000.

Good, I. J. & Card, W. I. (1971). The diagnostic process with reference to errors. *Meth. Inform. Med.* **10**, 176–88.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo Methods*. London: Methuen.

Lauder, I. J. (1978). Computational problems in predictive diagnosis. *Compstat 1978*, 186–92.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. *Proc. 4th Berkeley Symp.* **1**, 453–68.

Smyllie, H. C., Blendis, L. M. & Armitage, P. (1965). The observer disagreement in physical signs of the respiratory system. *The Lancet* **2**, 412–5.

[*Received December* 1978. *Revised May* 1979]

AITCHISON, J. (1979)

Calibration and assay from imprecise data

Reprinted from *Bull. Inst. Int. Statist.* 48, 4, 9-12

CALIBRATION AND ASSAY FROM IMPRECISE DATA

J. AITCHISON

University of Hong Kong

Summary

Calibration and assay involve the assessment of some unknown index of a specimen, not by direct measurement, but by inference from a comparison of the value of some related indicant of the specimen with the indicants of a training set of standards with known indices. The assumptions of most calibration models, namely that the indices and indicants of the training set and the indicant of the specimen are precisely determined, are often unrealistic. This paper indicates how such imprecisions can be incorporated into model-building and so provides a means of investigating the effects of imprecision in calibration and assay.

1. Calibration and assay with precise data

In a calibration or assay problem interest is in inferring for a case (blood sample, foetus, archaeological specimen, nuclear explosion) the value of some quantitative characteristic or index (plasma concentration of an antibiotic, length of pregnancy, age, nuclear yield). For reasons of cost, inconvenience, the time involved or even the destructiveness of the method of measurement we often hesitate to measure the index directly and would rather infer the index by measuring some related aspect, supplementary measure, or indicant (clearance diameter of droplet applied to an infected medium, crown-rump length of foetus in a sonar picture, radiocarbon count, magnitude of earth wave). The way in which an indicant x is related to the index t is seldom mathematically or even statistically known and usually has itself to be inferred from a calibration experiment, consisting of 'standards' or calibrative training set

$$D = \{(t_i, x_i) : i = 1, \ldots, n\} \qquad (1.1)$$

of n cases of known indices $t_i \in T$, the set of possible indices, and known indicants $x_i \in X_i$, the sets of possible indicants $(i = 1, \ldots, n)$. The simplest problem of calibration is then to make an inference about the unknown index u of a new case for which only the indicant y has been determined.

For the usual form of designed calibration experiment, which provides information only on the conditional distribution of indicant for given index, we adopt the assumptions and approach of Aitchison and Dunsmore (1975, §10.3) with inferential aim the

provision of a realistic calibrative density function $p(u|y,D)$ $(u \in T)$. This involves the adoption of parametric forms $p(x_i|t_i,\theta)$ $(i = 1,\ldots,n)$ and $p(y|u,\theta)$, where $\theta \in \Theta$, for the appropriate conditional density functions, with (y,D) the data and (u,θ) the parameter, θ being a 'nuisance' parameter.

To provide a backcloth against which to study modelling for imprecisions in the measurement of (y,D) we first draw attention to four basic questions which must be answered on the route towards the calibrative density function when all the components (y,D) are measured with precision.

(a) Likelihood evaluation. Can we evaluate the likelihood

$$L(u,\theta|y,D) = p(y|u,\theta) \prod_{i=1}^{n} p(x_i|t_i,\theta) \qquad (2.1)$$

and its derivatives with reasonable ease?

(b) 'Nuisance' parameter assessment. Can we obtain a realistic assessment of $p(\theta|D)$, assigning relative plausibilities to the possible values of the nuisance parameter θ from the data D of the training set?

(c) Construction of the predictive density function. Can we evaluate the 'predictive' density function (Aitchison and Dunsmore, 1975, §2.3):

$$p(y|u,D) = \int_{\Theta} p(y|u,\theta)\, p(\theta|D)\, d\theta? \qquad (2.2)$$

(d) Inversion of predictive to calibrative density function. For a given prior density function $p(u)$ on T can we evaluate the integral denominator in the application of Bayes's formula:

$$p(u|y,D) = p(u)p(y|u,\theta) \Big/ \int_{T} p(u)p(y|u,\theta)\, du?$$

For the case of precise (y,D), where the conditional distributions follow normal linear forms

$$p(x_i|t_i,\theta) = N(\alpha+\beta t_i,\sigma^2),\ p(y|u,\theta) = N(\alpha+\beta u,\sigma^2), \qquad (2.3)$$

Bayesian analysis leads straightforwardly to a predictive density function of general Student form $ST(K,A+Bu,C+Du+Eu^2)$ where K,A,B, C,D,E depend on (y,D). With precision of measurement therefore it is only in part (d) that any computational problem arises, and there only a simple quadrature is required to evaluate the denominator.

2. Calibration modelling for imprecise data

Imprecisions in the measurements of the data (y,D) of calibration and assay problems are common but are seldom recognised in

statistical modelling. Since the effects of analogous imprecisions in the structurally similar problem of statistical diagnosis, in indices by Aitchison and Begg (1976) and in indicants by Aitchison (1977, 1979), Aitchison and Lauder (1979), can be substantial it seems advisable to recognise the existence of any imprecisions in modelling for calibration. In the use of the indicant crown-rump length of foetus measured by sonar-screening to infer as index the age or 'menstrual maturity' of a new case of pregnancy imprecision in the indices of the training set may clearly arise due to inaccurate recording or faulty memory. Again, when the indicant itself, both in the training set and in the new case, can be measured only by another calibration or assay process then the recorded indicant is subject to imprecision, and we have what could be termed a problem of 'calibrating the calibrated'. Situations with imprecisions in all the components of the data (y,D) may arise, and so we attempt to model for all these eventualities.

First we suppose that imprecision in the true index u_i of a training case for which the recorded index is t_i may be expressed as a density function $p(u_i|t_i)$ on the index set T. For example, if t_i is an estimate of u_i with standard error s_i it may be a reasonable device to take $p(u_i|t_i) = N(t_i,s_i^2)$. Similarly if the measured indicants x_i $(i = 1,\ldots,n)$ and y are imprecise indicants of the true indicants, say w_i $(i = 1,\ldots,n)$ and w, we may be able to express this imprecision in terms of known conditional density functions

$$p(x_i|w_i) \quad (i = 1,\ldots,n), \quad p(y|w). \tag{2.4}$$

If the parametric models for the conditional density functions of true indicants on true indices are

$$p(w_i|u_i,\theta) \quad (i = 1,\ldots,n), \; p(y|u,\theta) \tag{2.5}$$

then we arrive at models in terms of the observable data (y,D) as

$$p(x_i|t_i,\theta) = \int_{W_i}\int_{V_i} p(x_i|w_i)p(w_i|u_i,\theta)p(u_i|t_i)\,dw_i\,du_i, \tag{2.6}$$

$$p(y|u,\theta) = \int_{W} p(y|w)p(w|u,\theta)\,dw, \tag{2.7}$$

and then the problems (a) - (d) have to be faced with these more complex parametric models.

With normal linear regression forms for all the conditional density functions it is easy to show that $p(x_i|t_i,\theta)$ and $p(y|u,\theta)$ take $N(\alpha+\beta t_i, A_i+B_i\beta^2+\sigma^2)$ and $N(\alpha+\beta u, B\beta^2+\sigma^2)$ forms with A_i, B_i and and B known. For precise data $A_i = B_i = B = 0$.

Thus, with known imprecision, problem (a) of likelihood eval-

uation presents no difficulty. For the resolution of problem (b) resort has usually to be made to asymptotic Bayesian maximum-likelihood theory, taking $p(\theta|D)$ as $N\{\hat{\theta}, V(\hat{\theta})\}$, where $\hat{\theta}$ is the maximum-likelihood estimate and $V(\hat{\theta})$ the usual estimate of the covariance matrix of the maximum-likelihood estimator. Problem (c) then involves the evaluation of a triple integral. Although this can be readily reduced to a single integral by a straightforward integration out with respect to α and β the resulting integral requires numerical integration. Since there remains yet a further integration step at problem (d) it is worth considering an approximation which avoids numerical integration at this stage. An appropriate approximation which retains the effects of all the sources of imprecision may be $p(x|u,D) \sim N\{a(u), b(u)\}$, where

$$a(u) = E(x|u,D) = \hat{\alpha} + \hat{\beta},$$

$$b(u) = V(x|u,D) = \hat{\sigma}^2 + B\hat{\beta}^2 + V(\hat{\sigma}) + BV(\hat{\beta}) + V(\hat{\alpha}) + 2C(\hat{\alpha},\hat{\beta})u + V(\hat{\beta})u^2.$$

Problem (d) then involves, as for precise data, the only and simple numerical quadrature in the stages leading to the calibrative density function.

The procedures developed here are currently being applied to a number of calibration problems involving imprecision, and the results will be reported elsewhere.

Résumé

Dans des problèmes de calibrage et d'essai il faut évaluer l'indice inconnu d'un specimen au moyen d'une comparaison d'un indiquant du spécimen avec les indiquants et les indices connus d'une ensemble d'instruction. Dans cet article on indique une méthode d'incorporer dans des modèles de calibrage et d'essai des imprécisions qui se présentent souvent dans ces indices et ces indiquants, et ainsi fournit un moyen d'étudier les effects d'imprécision.

References

AITCHISON, J. (1977). A calibration problem in statistical diagnosis: The system transfer problem. Biometrika 64, 461-72.

AITCHISON, J. (1979). A calibration problem in statistical diagnosis: The clinic amalgamation problem. Biometrika 66, to appear.

AITCHISON, J. & BEGG, C.B. (1976). Statistical diagnosis when the cases are not classified with certainty. Biometrika 63, 1-12.

AITCHISON, J. & DUNSMORE, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press.

AITCHISON, J. & LAUDER, I.J. (1979). Statistical diagnosis from imprecise data. Biometrika 66, to appear.

13 LOGISTIC NORMAL MODELS

13.1 Background

Some statisticians would argue strongly that all the worthwhile and tractable parametric classes of models have already been discovered and almost exhaustively developed. It is perhaps appropriate therefore to close this record, which started with a review and developments of the centenarian lognormal class, with a report on a new parametric class of models, the logistic-normal models. The emergence of this class has an interesting background. The one-dimensional version is a special case of the Johnson (1949) four-parameter lognormal distribution. A few uses of higher-dimensional versions are implicit in Bayesian multinomial analysis (Lindley, 1964; Swe, 1964; Bloch and Watson, 1967; Leonard 1973), in biometric shape-and-size studies (Mosimann, 1975a,b) and in the reconciliation of subjective probability assessments (Lindley, Tversky and Brown, 1979). The first traceable explicit definition of the class appears as a tentative suggestion in Johnson and Kotz (1972, p.20) where the idea is ascribed to Obenchain in a personal communication, but there seems to have been no subsequent development of the idea. Aitchison and Begg (16:1976) give independently a formal definition, recognising the class as a tractable substitute for the Dirichlet distribution for their particular application. Faced with a number of consultative problems involving compositional data, that is vectors whose components sum to unity, Aitchison and Shen (22:1980) realised the potential for statistical modelling of this class of models and first used the natural nomenclature of

logistic-normal.

The importance of this class of models is that it provides for the first time a rich class of distributions on the simplex to replace the very highly structured Dirichlet class. For example, Darroch and James (1974), in pointing to the almost-independence structure of the Dirichlet class, regret the fact that there seems to be a scarcity of distributions capable of describing situations which have any real structure of association. The logistic-normal class removes this scarcity.

Distributions over the simplex are fundamental to the study of compositional data, such as the chemical compositions of rocks in petrology, the composition of sediments, pollen compositions in palaeoecology, steroid compositions in blood sample analysis; and in the study of probabilistic data such as in the diagnostic statements of Aitchison and Begg (16:1976) and in inferential statements in studies of subjective performance in inferential tasks, as in Aitchison and Kay (14:1975), Aitchison and Shen (22:1980) and Aitchison (1980b).

13.2 Definition

Let R^d denote d-dimensional real space, P^d the positive orthant of R^d and S^d the d-dimensional positive simplex defined by

$$S^d = \{x \in P^d : x_1 + \ldots + x_d < 1\}.$$

Suppose that y follows a multinormal distribution $N_d(\mu,\Sigma)$ over R^d. The exponential transformation from R^d to P^d, namely $z = \exp(y)$, or its inverse the logarithmic transformation $y = \log(z)$, is the familiar device for defining a corresponding lognormal distribution with z distributed as $\Lambda_d(\mu,\Sigma)$, say. In a similar way the logistic

transformation from R^d to S^d, or its inverse logratio transformation:

$$x = \exp(y)/\{(1 + \sum_{j=1}^{d} \exp(y_j)\}, \quad y = \log(x/x_{d+1}),$$

where

$$x_{d+1} = 1 - \sum_{j=1}^{d} x_j ,$$

can be used to define a logistic-normal distribution over S^d with density function

$$|2\pi\Sigma|^{-\frac{1}{2}} \left(\prod_{j=1}^{d+1} x_j \right)^{-1} \exp[-\tfrac{1}{2}\{\log(x/x_{d+1})-\mu\}^T \Sigma^{-1} \{\log(x/x_{d+1})-\mu\}] \quad (x \in S^d).$$

13.3 Properties

Aitchison and Shen (1980) first set out to encourage the use of the logistic-normal class by enumerating some of its very attractive properties.

1. In situations where the proportions in the *composition* in S^d can be regarded as arising from actual quantities of a (d+1)-dimensional vector or *basis*, there is a simple relationship between logistic-normal compositions and lognormal bases.
2. Moment properties are easily derived.
3. There are interesting class-preserving transformations, including the permutation transformation.
4. The distribution of a subcomposition is readily related to the full compositional distribution.
5. The conditional distribution of a subcomposition, given another subcomposition, is also readily obtained.
6. The close relationship to the multivariate normal distribution ensures simple statistical analysis. In particular statistical

prediction analysis is available, and gives rise to other distributions on the simplex, such as the logistic-Student distribution.

7. The relationship of the logistic-normal class to the Dirichlet $D_d(\alpha)$ class can be explored. In particular the following aspects are shown.

   (a) The $L_d(\mu,\Sigma)$ closest to $D_d(\alpha)$ in the sense that the Kullback-Liebler (1951) measure of directed divergence of $L_d(\mu,\Sigma)$ from $D_d(\alpha)$ is least, has

$$\mu_i = \delta(\alpha_i) - \delta(\alpha_{d+1})$$

$$\sigma_{ii} = \varepsilon(\alpha_i) + \varepsilon(\alpha_{d+1}) \qquad (i = 1,\ldots,d),$$

$$\sigma_{ij} = \varepsilon(\alpha_{d+1}) \qquad (i \neq j),$$

   where δ and ε are the digamma and trigamma functions.

   (b) A preliminary study of the extent of this closeness suggests that in practice any Dirichlet distribution can be effectively replaced by a logistic-normal.

   (c) It is shown that the above result gives a global version of an approximation used by Bloch and Watson (1967) in a study of logarithmic contrasts in the analysis of contingency tables.

13.4 Applications

As indicated in §13.3 the direct relationship of the logistic-normal class to the normal class opens up all the available statistical analysis associated with normal theory. In particular, Aitchison and Shen (22:1980) provide a selection of possible applications.

1. The application to the study of log contrasts in the analysis of multinomial distributions and contingency tables, as mentioned above.
2. Applications to the analysis of compositional data, including a discriminant problem involving a new predictive density function, the logistic Student-Siegel, hypothesis testing problems involving two logistic-normal distributions, the evaluation of atypicality indices and conditional compositions through the use of the new logistic-Student predictive distribution.
3. Whereas Aitchison and Kay (15:1975) and Kay (1976) analyse subjective performance in inferential tasks in terms of real-valued constructs of the diagnostic statements, such as inference discrepancy and information gain index, it is now possible to analyse the inferential statements as probabilistic vectors. An illustrative example compares the performance of two groups of students, one group being unfamiliar with, the other familiar with, Bayes's formula.
4. An application to logistic discriminant analysis provides a picture of the unreliability of the discriminant process.

Two further specific applications are considered in Aitchison (23:1980a, 24:1980b).

There is a now considerable literature on the spurious correlations that may arise in proportions comprising a composition, even although the basis of quantitative measurements has statistically independent components, the paper by Pearson (1897) probably being the earliest. Over the last two decades there has been quite intensive study (Chayes, 1960, 1962; Chayes and Kruskal, 1966; Mosimann, 1962, 1963; Darroch, 1969; Darroch and Ratcliff, 1970,

1978; Darroch and James, 1974; Bartlett and Darroch, 1978), mostly in the area of geological applications, into the nature of these 'null correlations' arising from an independent basis, and attempts to construct tests of departure from these null correlations. Unfortunately these attempts are unsatisfactory since the distributions of the test statistics are not known (Mosimann, 1962, p.81; Chayes and Kruskal, 1966, p.696), the tests of null correlations are carried out separately for each pair of proportions rather than as one overall test (Chayes and Kruskal, 1966, p.696), and even when the tests detect non-null correlations it is by no means safe (Miesch, 1969) to conclude that the corresponding quantities in the basis are uncorrelated.

Aitchison (21:1980) suggests that the unsatisfactory features of the above theories can be largely remedied by realising that there is a simple and exact relationship between the covariance structure of the logarithms of the basis components and the covariance structure of a composition. In particular, if the basis has independent components with variances of the logarithms $\omega_1,\ldots,\omega_{d+1}$ then the covariance matrix of the logratio proportions is

$$\mathrm{diag}(\omega_1,\ldots,\omega_d) + \omega_{d+1}\, U_d,$$

where U_d is a $d \times d$ matrix of units. It is then possible to construct from standard statistical theory using the logistic-normal distributions as the class of composition models a reasonable test of whether compositional data conform to this structure, that is whether the hypothesis of basis independence is tenable. Complete details of this test are provided and applications are given to fossil pollen counts, volcanic rock compositions in Taupo and Skye,

and sediment variability.

It is emphasised that, while the logistic normal provides a satisfactory means of testing for basis independence its great strength lies in its flexibility for the investigation of dependence structures.

When a composition $x \in S^d$ is formed from a basis of d+1 measurements $y_1,\ldots,y_{d+1}$, as

$$x_i = y_i/(y_1 + \ldots + y_{d+1}) \quad (i = 1,\ldots,d),$$

the question of whether the composition is independent of additive size $z = y_1 + \ldots + y_{d+1}$ is an important one. This independence is the *additive isometry* of Mosimann (1975) in his shape and size studies in biometric allometry, and the *proportional invariance* of Darroch and James (1974) in discussing the validity of compositional analysis. So far no satisfactory test of such independence seems to have been devised. Aitchison (24:1980) provides an explanation of why this lack has persisted and proceeds to remedy it.

The explanation suggests that Mosimann's (1975b) characterisation property of the lognormal distribution, that any lognormal basis with additive isometry is necessarily degenerate, has seemed to rule out the use of the most tractable lognormal class from considerations of modelling in compositional data analysis. Aitchison (24:1980) points out that there is no great merit in insisting on a lognormal *basis* and that it is quite possible to have a logistic-normal composition, a lognormal size *and* additive isometry. Moreover, the pattern of joint variability of composition and additive size can be so modelled that additive

isometry or proportional invariance becomes a parametric hypothesis testable through standard hypothesis testing methodology. The test is illustrated by application to a problem of additive isometry in measurements of heart shape and size, and to a problem of proportional invariance in the steroid metabolite composition in urinary excretion.

AITCHISON, J. and SHEN, S.M. (1980)

Logistic-normal distributions: some properties and uses

To appear in *Biometrika* 67

Reprinted from page proofs

Biometrika (1980), **67**
Printed in Great Britain

Logistic-normal distributions: Some properties and uses

BY J. AITCHISON AND S. M. SHEN
Department of Statistics, University of Hong Kong

SUMMARY

The logistic transformation applied to a d-dimensional normal distribution produces a distribution over the d-dimensional simplex which can sensibly be termed a logistic-normal distribution. Such distributions, implicitly used in a number of recent applications, are here given a formal identity and some useful properties are recorded. A main aim is to extend the area of application from the restricted role as a substitute for the Dirichlet conjugate prior class in the analysis of multinomial and contingency table data to the direct statistical description and analysis of compositional and probabilistic data.

Some key words: Compositional data; Directed divergence measure; Dirichlet distribution; Logistic discrimination; Logistic-normal distribution; Log normal distribution; Multiple contingency table; Probabilistic data.

1. DEFINITION

Most statisticians if asked to list the distributions they know over the unit interval $(0, 1)$ and its generalization, the d-dimensional simplex S^d, would start and end with the class of beta distributions and their higher-dimensional counterpart, the Dirichlet distributions. Another useful and richer such class, the logistic-normal distributions, has arisen usually by implication in a few widely differing applications:

(i) for the case $d = 1$ as a Johnson (1949) four-parameter log normal distribution with the two range parameters determining the interval $(0, 1)$;

(ii) in the Bayesian analysis of multinomial and contingency table data in the use of normal approximations to log contrasts by Lindley (1964), by C. Swe in a Liverpool Ph.D. dissertation, and by Bloch & Watson (1967); and as the first stage in the construction of exchangeable prior distributions by Leonard (1973);

(iii) in studies of size and shape in biological allometry, for example by Mosimann (1975), as the distribution of ratios of log normally distributed measurements;

(iv) in statistical diagnosis where classification of the basic cases is subject to uncertainty, as discussed by Aitchison & Begg (1976), who provide an explicit definition of the class of logistic-normal distributions;

(v) in the reconciliation of subjective probability assessments, where Lindley, Tversky & Brown (1979) use normal log-odds models to describe assessments.

The class has, however, gained no clear identity. Our purpose here is to provide such an identity, to describe enough of its properties and a sufficient variety of new applications to encourage further exploration of what, in our view, is a promising tool of statistical analysis.

Let R^d denote d-dimensional real space, P^d the positive orthant of R^d and S^d the d-dimensional positive simplex defined by

$$S^d = \{u \in P^d : u_1 + \ldots + u_d < 1\}.$$

For any d-vector u and any real-valued function f, let $f(u)$ denote the d-vector with ith component $f(u_i)$ $(i = 1, \ldots, d)$. Suppose that v follows a multinormal distribution $N_d(\mu, \Sigma)$

over R^d. The exponential transformation from R^d to P^d, namely $w = e^v$, or its inverse the logarithmic transformation $v = \log w$, is the familiar device for defining a corresponding log normal distribution with w distributed as $\Lambda_d(\mu, \Sigma)$, say. In a similar way the logistic transformation from R^d to S^d, or its inverse log ratio transformation:

$$u = e^v \Big/ \left(1 + \sum_{j=1}^{d} e^{v_j}\right), \quad v = \log(u/u_{d+1}), \tag{1·1}$$

where

$$u_{d+1} = 1 - \sum_{j=1}^{d} u_j, \tag{1·2}$$

can be used to define a logistic-normal distribution over S^d, and we can say briefly that u is $L_d(\mu, \Sigma)$. The density function of $L_d(\mu, \Sigma)$ is then

$$|2\pi\Sigma|^{-\frac{1}{2}} \left(\prod_{j=1}^{d+1} u_j\right)^{-1} \exp\left[-\tfrac{1}{2}\{\log(u/u_{d+1}) - \mu\}^{\mathrm{T}} \Sigma^{-1} \{\log(u/u_{d+1}) - \mu\}\right] \quad (u \in S^d). \tag{1·3}$$

Note that the density function is defined on the strictly positive simplex. This is necessary because of the logarithmic transformation involved.

We first look at properties of the logistic-normal distribution in § 2, compare it in § 3 with its competitor, the Dirichlet distribution, describe applications in § 4, and finally in § 5 draw some conclusions and point the way to further research and applications.

2. PROPERTIES

2.1. *Compositions*

Throughout this section the d-dimensional random vectors u and v have $L_d(\mu, \Sigma)$ and $N_d(\mu, \Sigma)$ distributions. Most of the properties of logistic-normal distributions derive from corresponding properties of multinormal distributions but usually require adjustment to provide useful practical results. For example, normality of the marginal distribution of $(v_1, \ldots, v_{c+1})$ over R^c does not provide a simple result about the distribution of $(u_1, \ldots, u_{c+1})$ over S^{c+1}, that is about $u_1, \ldots, u_{c+1}$ and the residual $1 - u_1 - \ldots - u_{c+1}$, but rather about the relative structure within the subvector $(u_1, \ldots, u_{c+1})$.

This and other properties are most simply expressed in terms of the concept of the composition and subcompositions of a vector. The composition of any positive $(d+1)$-vector w is the d-vector u defined by $u_i = w_i/(w_1 + \ldots + w_{d+1})$ $(i = 1, \ldots, d)$, is written $C(w)$, and is an element of S^d. The composition of any subvector of u, such as $(u_1, \ldots, u_{c+1})$ is then a subcomposition of u or of w, and is an element of S^c.

Proofs of the following properties are straightforward and are therefore omitted.

2.2. *The composition of a log normal vector*

If w is $\Lambda_{d+1}(\xi, \Omega)$ then $C(w)$ is $L_d(A\xi, A\Omega A^{\mathrm{T}})$, where the $d \times (d+1)$ matrix $A = [I_d - e_d]$, I_d is the unit matrix of order d and e_d is a d-vector with unit components.

This result is of particular interest in § 4·2 where we study problems concerning the analysis of compositional data.

2.3. *Moment properties*

Although moments of all positive orders $E(u_j^a)$ $(a > 0)$ and the geometric moment $\exp\{E[\log u_j]\}$ exist the integral expressions for them are not reducible to any simple form.

This is no great loss since interest in practice is often more naturally in the ratios u_j/u_k or their logarithms. From normal-log normal theory, with σ_{jk} denoting the (j,k)th element of Σ and with the convention that $\mu_{d+1} = 0$ and $\sigma_{j,d+1} = 0$ $(j = 1, \ldots, d+1)$, we have that

$$E\{\log(u_j/u_k)\} = \mu_j - \mu_k, \quad \operatorname{cov}\{\log(u_j/u_k), \log(u_l/u_m)\} = \sigma_{jl} + \sigma_{km} - \sigma_{jm} - \sigma_{kl},$$

$$E(u_j/u_k) = \exp\{\mu_j - \mu_k + \tfrac{1}{2}(\sigma_{jj} - 2\sigma_{jk} + \sigma_{kk})\},$$

$$\operatorname{cov}(u_j/u_k, u_l/u_m) = E(u_j/u_k)\,E(u_l/u_m)\{\exp(\sigma_{jl} + \sigma_{km} - \sigma_{jm} - \sigma_{kl}) - 1\}.$$

2.4. *Class-preserving properties*

The well-known linear transformation property of multinormal distributions, that if v is $N_d(\mu, \Sigma)$ and B is a $c \times d$ matrix then Bv is $N_c(B\mu, B\Sigma B^{\mathrm{T}})$, has the following counterpart in logistic-normal theory. If u is $L_d(\mu, \Sigma)$ then the c-vector t, defined by

$$t_i = \prod_{j=1}^{d} (u_j/u_{d+1})^{b_{ij}} \left\{1 + \sum_{i=1}^{c} \prod_{j=1}^{d} (u_j/u_{d+1})^{b_{ij}}\right\}^{-1} \quad (i = 1, \ldots, c), \tag{2·1}$$

is $L_c(B\mu, B\Sigma B^{\mathrm{T}})$. Two special cases of this property are of particular importance.

We first consider the permutation property. In our definition of the logistic-normal distribution over S^d, u_{d+1} was the common divisor in all the ratios of the transformation (1.1). A first application of (2.1) with $c = d$, $b_{ii} = 1$ $(i \neq h)$, $b_{ih} = -1$ $(i = 1, \ldots, d)$, $b_{ij} = 0$ otherwise shows that the d-vector t defined by $t_i = u_i$ $(i \neq h)$, $t_h = u_{d+1}$, $t_{d+1} = u_h$ is also of L_d form, with ratio denominator t_{d+1} now effectively the original u_h. This class preservation property is reassuring for any work in the simplex, where obviously it is a matter of no consequence which d of the $d+1$ positive quantities $u_1, \ldots, u_{d+1}$ are chosen to define the simplex of interest.

A second application of (2·1), with B the $d \times d$ permutation matrix associated with the permutation $(1, \ldots, d) \to (j_1, \ldots, j_d)$ so that $b_{ij_i} = 1$ $(i = 1, \ldots, d)$, $b_{ij} = 0$ otherwise, gives $t_i = u_{j_i}$ $(i = 1, \ldots, d)$. This shows that the logistic-normal form is preserved under a permutation of $u_1, \ldots, u_d$.

Combination of these two preceding results establishes that the logistic-normal class of distributions is closed under the group of permutations of the components $u_1, \ldots, u_d, u_{d+1}$. This is particularly reassuring in many statistical investigations of vector data, where we hope that the analysis is invariant with respect to the ordering of the vector components.

For the subcomposition property, a useful counterpart of the multinormal marginal property is obtained from (2·1) with B the $c \times d$ matrix defined by $b_{ii} = 1$,

$$b_{i,c+1} = -1 \ (i = 1, \ldots, c),$$

$b_{ij} = 0$ otherwise. Then $t_i = u_i/(u_1 + \ldots + u_{c+1})$ $(i = 1, \ldots, c)$ and the property is simply that the subcomposition $C(u_1, \ldots, u_{c+1})$ is $L_c(B\mu, B\Sigma B^{\mathrm{T}})$.

2.5. *The conditional subcomposition property*

The multinormal conditional property can be adjusted to provide a useful conditional distribution property for subcompositions. Suppose that the subcomposition $C(u_{c+1}, \ldots, u_{d+1})$ is known, expressed most conveniently for our purposes in terms of specified values r_i of u_{c+i}/u_{d+1} $(i = 1, \ldots, d-c)$. The conditional distribution of the subcomposition $C(u_1, \ldots, u_{c+1})$ given the above subcomposition $C(u_{c+1}, \ldots, u_{d+1})$ is then

$$L_c\{\mu_1 - e_c \log r_1 + \Sigma_{12}\Sigma_{22}^{-1}(\log r - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\},$$

where (μ_1, μ_2) is the $(c, d-c)$ partition of μ and $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}$ are the obvious submatrices in the

corresponding partition of Σ. From this the conditional distribution of $C(u_1, \ldots, u_c)$ given $C(u_{c+1}, \ldots, u_{d+1})$ can easily be obtained from the subcomposition property of § 2·4.

2·6. *Statistical properties*

For the analysis of compositional data, such as independent vectors $u^{(1)}, \ldots, u^{(n)}$ with each $u^{(i)} \in S^d$ $(i = 1, \ldots, n)$, the class of logistic-normal distributions provides, through its close relationship to the multinormal class, a ready means of tractable statistical analysis. In addition to simple estimation and hypothesis testing of the μ and Σ parameters, tests of logistic-normality, linear modelling of $E(u^{(i)})$ $(i = 1, \ldots, n)$ to take account of experimental design and possible concomitant factors, and all the special multivariate techniques such as discriminant analysis, Bayesian statistical analysis is directly available through the normal-Wishart class of conjugate prior distributions for μ and $\tau = \Sigma^{-1}$. In particular, predictive density functions are easily derived in terms of new and easily defined classes of distributions such as logistic-Student distributions over S^d. We shall see applications of such predictive distributions in § 4.

3. COMPARISON WITH THE DIRICHLET CLASS

3·1. *Closeness*

No study of a proposed class of distributions over the simplex can fail to make comparisons with the well established Dirichlet class, with typical distribution $D_d(\alpha_1, \ldots, \alpha_{d+1})$ or $D_d(\alpha)$ defined by a density function proportional to

$$\prod_{i=1}^{d+1} u_i^{\alpha_i - 1},$$

where $u_{d+1} = 1 - u_1 - \ldots - u_d$ as before.

Aitchison & Begg (1976) suggest that the greater richness of the L_d class with its $\frac{1}{2}d(d+3)$ parameters compared with only $d+1$ for the D_d class, may allow any Dirichlet distribution to be closely approximated by a suitably chosen logistic normal distribution. Their suggestion can be investigated more quantitatively by means of the Kullback & Liebler (1951) measure of directed divergence of a density function q from a target density function p:

$$I(p, q) = \int_{S^d} p(u) \log \frac{p(u)}{q(u)} du.$$

For $p(u)$ of $D_d(\alpha)$ form the closest $q(u)$ of $L_d(\mu, \Sigma)$ form in the sense that $I(p, q)$ is minimized is given by

$$\mu_i = \delta(\alpha_i) - \delta(\alpha_{d+1}), \quad \sigma_{ii} = \varepsilon(\alpha_i) + \varepsilon(\alpha_{d+1}) \quad (i = 1, \ldots, d), \quad \sigma_{ij} = \varepsilon(\alpha_{d+1}) \quad (i \neq j), \qquad (3{\cdot}1)$$

where $\delta(x) = \Gamma'(x)/\Gamma(x)$ and $\varepsilon(x) = \delta'(x)$ are the digamma and trigamma functions.

For $D_d(\alpha)$ with $d = 1, 2, 3$, and with components of α in the range 5 to 100, the minimized divergences range from 2×10^{-6} for $\alpha = (5, 5)$ to 5×10^{-2} for $\alpha = (5, 5, 5, 100)$.

Some indication of the degree of closeness of these approximations can be provided by directed divergences for more familiar situations. For example, the directed divergence of a $N(\lambda, 1)$ from the $N(0, 1)$ distribution also ranges from 2×10^{-6} to 5×10^{-2} as λ ranges from 0·002 to 0·316.

We can also try to judge success for given $D_d(\alpha)$ by finding a neighbouring $D_d(\beta)$ distribution with the same directed divergence as the minimized logistic-normal divergence. Confining attention to neighbouring distributions with α differing from β in a single component we have the following results. When the components of α are equal the increase in a

single α_i never exceeds 0·6. When the components are unequal, the more asymmetrical the component values are, the greater, roughly speaking, is the increase in these single component values, the greatest increases being 0·7 in 5, 1·4 in 20 and 8·6 in 100, the last occurring for $d = 3$ and $\alpha = (5, 5, 5, 100)$.

Whether or not such results are acceptable approximations must clearly depend on the particular application. We shall see in §4·1 that minimized-divergence logistic-normal distributions are indeed close to an already widely accepted approximation to Dirichlet distributions.

3·2. *Comparison of properties*

Some comparison of the Dirichlet and logistic-normal classes with respect to the properties of §2 is required.

The Dirichlet composition property analogous to §2·2 relates a Dirichlet-distributed composition u to a $(d+1)$-vector with gamma-distributed components (Wilks, 1962, p. 179). The fact that the gamma components are independent and have equal scale parameters indicates that the components of a Dirichlet composition have a special and near-independence structure, with correlations between components arising solely from the division by the common Σw_i in the formation of the composition u. Thus Dirichlet distributions may be too simple to be realistic in the analysis of compositional data where the underlying w_i's are dependent.

The Dirichlet class has simple analogues to the class-preserving properties of §2·4. Note that the subcomposition property of §2·4 is not a result about $(u_1, \ldots, u_{c+1}, 1-u_1-\ldots-u_{c+1})$ but about the distribution of $u_1/(u_1+\ldots+u_{c+1}), \ldots, u_c/(u_1+\ldots+u_{c+1})$. In other words the logistic-normal subcomposition property is not a class-preserving property allowing addition of components of a composition. The Dirichlet does possess such a component-additive property (Wilks, 1962, p. 181): for example $(u_1, \ldots, u_{c+1}, 1-u_1-\ldots-u_{c+1})$ is $D_{c+1}(\alpha_1, \ldots, \alpha_{c+1}, \alpha_{c+2}+\ldots+\alpha_{d+1})$. This is, however, a direct consequence of the compositional relationship to independent gamma variables and so, as an advantage over the logistic-normal class, may be buying mathematical elegance at the price of realism. Moreover, if Dirichlet distributions are truly appropriate to an analysis involving additions of compositional data then the logistic-normal distributions that we mistakenly use may yet prove to be satisfactory understudies through the closeness property.

The Dirichlet counterpart of the conditional subcomposition property of §2·5 reinforces this caution in the use of Dirichlet distributions in the analysis of compositional data. For the Dirichlet distribution the conditional distribution of $C(u_1, \ldots, u_c)$ given $C(u_{c+1}, \ldots, u_{d+1})$ is the same as the unconditional distribution $C(u_1, \ldots, u_c)$. In other words in Dirichlet modelling $C(u_1, \ldots, u_c)$ and $C(u_{c+1}, \ldots, u_{d+1})$ are independent, a very strong assumption to impose, without investigation, on the nature of any compositional data.

Although the Dirichlet class may possess admirable qualities of mathematical tractability in its role as the conjugate prior class for the Bayesian analysis of multinomial and contingency table data, and as an essential tool in the determination of distribution-free statistical tolerance limits it has many disadvantages as a direct describer of patterns of variability. Maximum likelihood estimation of α, for example, requires solving equations involving digamma functions so that Newton–Raphson or some equivalent numerical method is required; and the distributional properties of the estimators must also be approximated. Moreover the absence of a class of conjugate priors makes the possibility of tractable Bayesian analysis and statistical prediction analysis remote.

4. SOME APPLICATIONS

4·1. *Bayesian analysis of contingency tables*

For the $(d+1)$-category multinomial distribution with category probabilities

$$\theta_i \ (i = 1, \ldots, d+1),$$

the conjugate class of distributions on the parameter space S^d is the Dirichlet class. When the multinomial distributions relate to contingency tables then interest is often concerned with contrasts, linear combinations of $\log \theta_i$ such as

$$k_h = \sum_{i=1}^{d+1} c_{hi} \log \theta_i$$

with $\Sigma_i c_{hi} = 0$. From the logistic-normal approximation (3·1) it follows immediately that k_h is approximately distributed as $N\{\Sigma_i c_{hi}\, \delta(\alpha_i), \Sigma_i c_{hi}^2\, \varepsilon(\alpha_i)\}$ and moreover that k_g and k_h are approximately distributed binormally with means and variances as determined above and with covariance $\Sigma_i c_{gi} c_{hi}\, \varepsilon(\alpha_i)$.

Bloch & Watson's (1967) approximation to such contrasts, essentially derived from a component by component choice of expressions for means and variances of the logarithm of gamma random variables, uses Stirling's approximation to the log gamma function and so could easily have led to a digamma-trigamma approximation coincident with our own, whose derivation is based on a global approximation to the Dirichlet distribution. Presumably their determination to arrive at approximate means and variances expressible in terms of logarithmic and reciprocal functions was motivated by a wish to avoid digamma and trigamma functions. Our global approximation provides overall support for the component-wise method used by Bloch & Watson. The directed divergence of their approximation from $D_d(\alpha)$ is greater than that of the minimizing logistic-normal distribution (3·1) by less than 0·1% when the components of α are all greater than 2, which will almost always be the case in applications.

4·2. *Analysis of compositional data*

There are many disciplines, for example, sedimentology, petrology, biochemistry, palaeoecology, where interest is in compositional data such as proportions of sand, silt, clay in sediments, of chemical constituents of rocks, of serum proteins in blood, of pollens of different species at different levels in sample borings. For illustrative purposes we here adapt a problem posed by McCammon (1975, p. 162) to demonstrate the simplicity of statistical analysis with logistic normal tools.

Figure 1 shows in terms of triangular coordinates the sand, silt, clay composition of 17 sediments, 7 of which are identified as nearshore, type I, and the remaining 10 as offshore, type II. Four new samples, all from the same site and hence of the same type, have been analysed and the problem is to assess this unknown type. We adopt $L_2(\mu_1, \Sigma_1)$ and $L_2(\mu_2, \Sigma_2)$ distributions for the nearshore and offshore data. The statistical problem is assumed to be the assessment of a reasonable factor for the conversion of prior odds to posterior odds for type. With such a small data set we adopt a predictive approach to the typing or diagnostic problem, for the advantages argued, for example, by Aitchison, Habbema & Kay (1977).

The unusual feature of this example, in contrast to the more familiar area of application of predictive diagnosis, namely medical diagnosis, is that for the new case we have four replicate observations. For the purposes of predictive diagnosis the predictive problem can then be condensed into obtaining a predictive density function for M and V, the mean

vector and matrix of corrected cross-products of the four, in general N, vectors of log ratios. The appropriate theory is summarized by Aitchison & Dunsmore (1975, Table 2·3) and leads to predictive distributions $p(M, V \mid D_i)$ for M and V of Student–Siegel type based on data D_i:

$$\mathrm{St\,Si}_2\left\{n_i - 1, m_i, \left(\frac{1}{n} + \frac{1}{N}\right) S_i;\ N - 1, (n_i - 1)\, S_i\right\},$$

where n_i, m_i and S_i are the number, the mean vector and covariance matrix of the log ratios, of vectors in the data set D_i for type i. In this assessment we have used the vague prior suggested by Aitchison & Dunsmore (1975) for the reasons given by Aitchison (1976). Straightforward computation then gives $p(M, V \mid D_1)/p(M, V \mid D_2) = 0{\cdot}19$ as the converting factor from prior odds to posterior odds on type I. Thus if type I and II are equally likely *a priori* our evidence leads to odds of 5 to 1 in favour of type II.

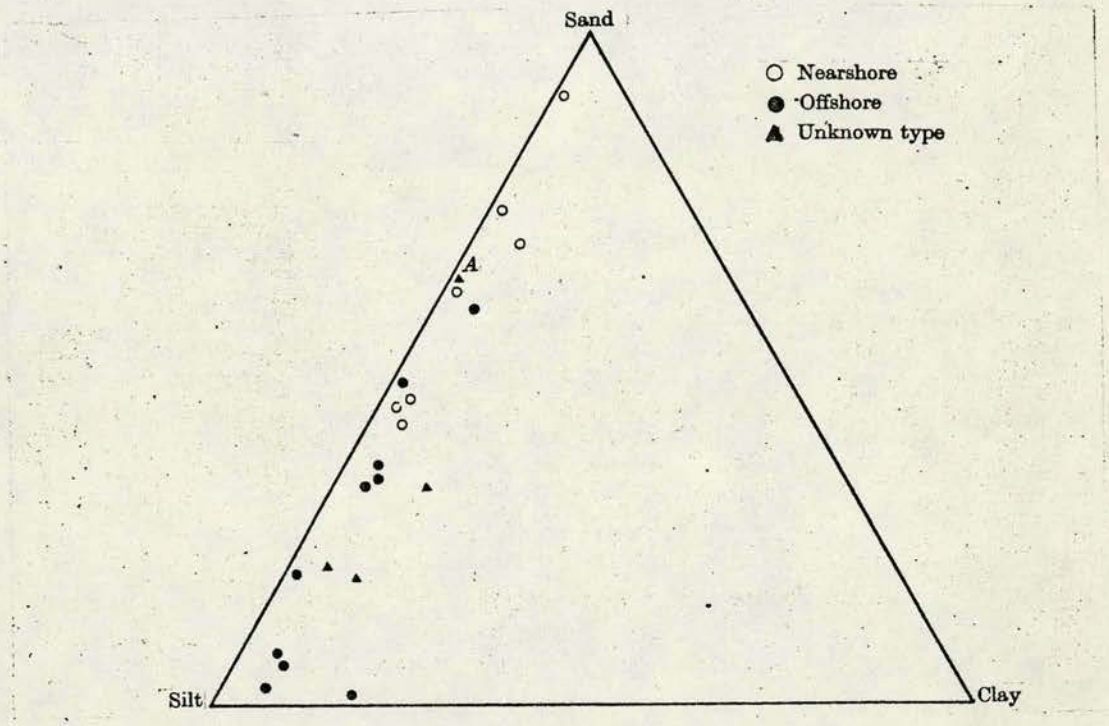

Fig. 1. Sand, silt, clay compositions of 17 sediment specimens of known type and four sediment samples of unknown type.

The predictive method can be contrasted with the estimative method (Aitchison, Habbema & Kay, 1977) which would simply replace the parameters $\mu_1, \Sigma_1, \mu_2, \Sigma_2$ by their estimates. This is equivalent to replacing the previous $p(M, V \mid D_i)$ by a normal-Wishart density function $\mathrm{NoWi}_2(m_i, S_i/N, N-1, S_i)$ as defined by Aitchison & Dunsmore (1975, Table 2·1), leading to the replacement of the odds of 5 to 1 by odds of approximately 80 to 1. Examination of Fig. 1 suggests that these latter odds are extravagant, illustrating the tendency of estimative methods to read too much into the data.

The proposition of the specimen labelled A in Fig. 1 raises the question of whether it is atypical of the identified offshore standards. To examine this its atypicality index, defined by Aitchison & Dunsmore (1975, p. 226) as the probability that a case has a higher predictive density than the case under scrutiny, may be evaluated from formula (11·20) of Aitchison & Dunsmore (1975). The atypicality index is only 0·62 and so the specimen can hardly be regarded as atypical.

This example can be further used to illustrate the nature of the conditioning property. Suppose that for an offshore specimen we wish to study the variability of the composition of (sand, clay) for a given silt/clay ratio. We can proceed as follows. First find the predictive distribution for a new complete vector based on D_2. This will be two-dimensional logistic-Student with 9 degrees of freedom. We can then easily derive the predictive conditional distribution of (sand, clay) for given value v of log (silt/clay) as logistic-Student with 9 degrees of freedom. In more familiar terms the conditional distribution of log (sand/clay) is

$$\mathrm{St}_1\{9, -3{\cdot}29+1{\cdot}66v, 2{\cdot}10+0{\cdot}422(v-2{\cdot}71)^2\}$$

in the notation of Aitchison & Dunsmore (1975, Table 2·2). Figure 2 shows the considerable differences in this logistic Student distribution for three silt/clay ratios spanning the range of silt/clay ratios observed in the specimens.

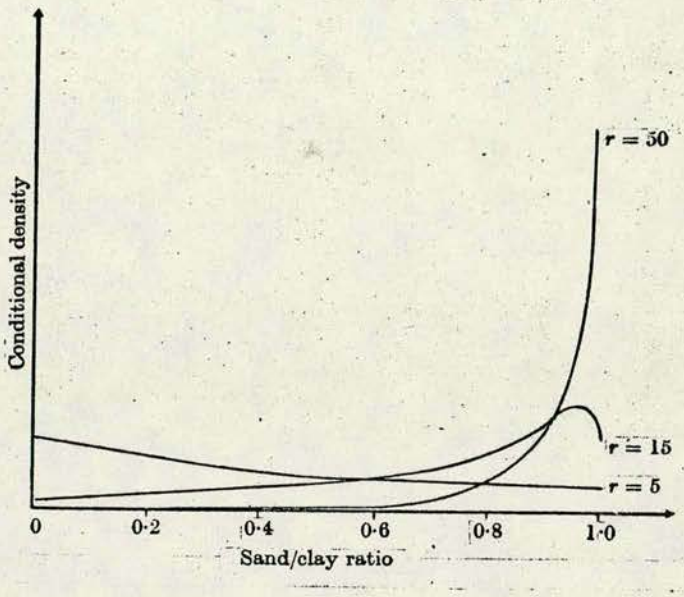

Fig. 2. Conditional density functions of (sand, clay) composition for given value r of silt/clay ratio.

4·3. *Analysis of probabilistic data*

Data in the form of probability vectors arise in a variety of applications such as answers to multiple-choice questions (de Finetti, 1972, p. 30) and in the study of subjective performance in inferential tasks (Taylor, Aitchison & McGirr, 1971). Provided that the probabilities are all positive we then have data in the form of vectors in S^d.

To obtain a simple illustration we presented 24 students with exactly the same diagnostic problem, the differential diagnosis of newmath syndrome (Aitchison, 1974), and asked them to assess subjectively the diagnostic probabilities they attached to each of three possible types. Conditions were identical for all students except that 12, randomely selected, performed the task before, and the remaining 12 after, they had encountered the appropriate statistical tool, Bayes's formula. The diagnostic assessments form two sets of probabilistic data, which can conveniently be presented in triangular coordinates in Fig. 3. A question of interest is then whether there is any significant difference in performance in the after and before groups. Adopting logistic-normal distributions $L_2(\mu_A, \Sigma_A)$ and $L_2(\mu_B, \Sigma_B)$ to describe the variability in the after and before data we can then test differences in terms of standard

multinormal tests (Anderson, 1958, Chapter 10) of $\mu_A = \mu_B$ and $\Sigma_A = \Sigma_B$. No significant differences are found.

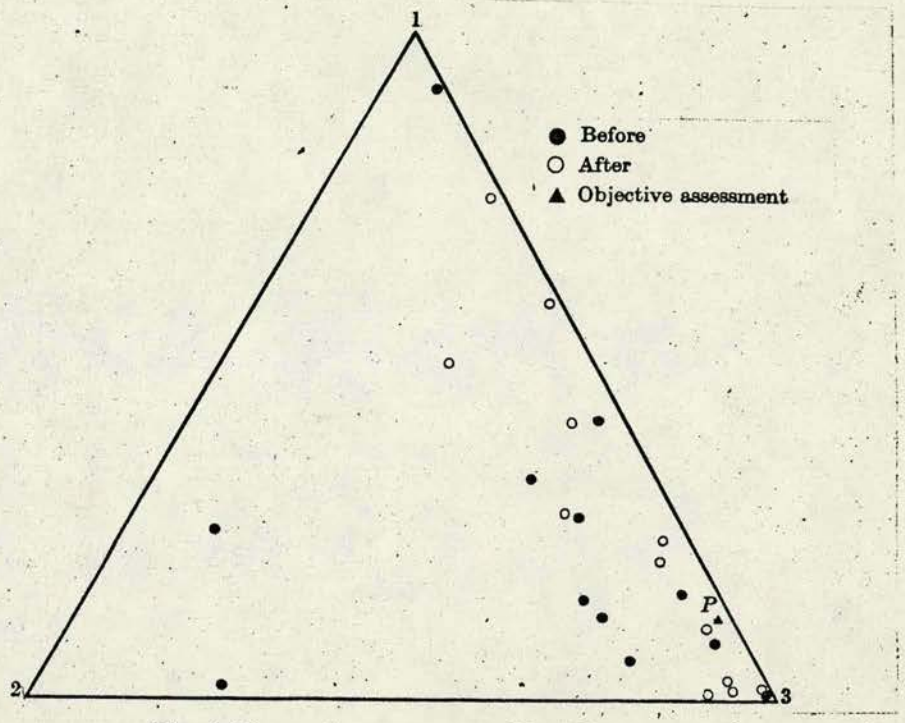

Fig. 3. Diagnostic assessments of 24 subjects.

This particular inference task has an objective answer $P = (0{\cdot}14, 0{\cdot}02, 0{\cdot}84)$ shown on Fig. 3. We can then investigate the extent to which the subjective inferences depart from this by testing the hypotheses $\mu_A = \mu$ and $\mu_B = \mu$, where $\mu = (-1{\cdot}79, -3{\cdot}74)$, the log ratio vector associated with P. Both these tests give significant departure from μ, at the 1% significance level for μ_A and at the 5% significance level for μ_B.

4·4. *An application in logistic discriminant analysis*

For simplicity we confine attention to discrimination between two types. In logistic discriminant analysis the probability that a case with given vector x of diagnostic features is of type t ($t = 1, 2$) is expressed in the parametric form

$$p(t = 1 \mid x, \beta) = 1 - p(t = 2 \mid x, \beta) = \exp(\beta^{\mathrm{T}} x)/\{1 + \exp(\beta^{\mathrm{T}} x)\},$$

where β denotes the parameter.

Standard practice here is to use the diagnostic training set $D = \{(t_i, x_i) : i = 1, \ldots, n\}$ of n cases of known types t_i and corresponding features vectors x_i first to estimate β by its maximum likelihood estimate $\hat{\beta}$. To produce diagnostic probabilities for a new case with feature vector x a common procedure is then simply to quote the estimative probabilities $p(t \mid x, \hat{\beta})$ or even to leave these in their transformed or log odds versions, commonly termed the scores $\hat{\beta}^{\mathrm{T}} x$. There is usually little attempt to quantify in a meaningful way the reliability of such a diagnostic assessment other than to make some comment about the possibilities of producing standard errors for the scores. One way of taking account of the unreliability of the estimation process in reaching diagnostic probabilities is through the predictive diagnostic device (Aitchison, Habbema & Kay, 1977) of weighting each possible assessment $p(t \mid x, \beta)$ by a

suitable posterior distribution $p(\beta|D)$ to obtain diagnostic probabilities $\int p(t|x,\beta)p(\beta|D)d\beta$, where the integral is over the range $\mathcal{B}$ which is the set of possible parameters β. Although this device does take account of the unreliability it is sometimes criticized because its presentation of a single set of diagnostic probabilities gives the impression that these are the diagnostic probabilities rather than the result of a weighting process. For more than two types this weighting process is probably the only realistic way of presenting a comprehensible and practically useful overall view. For two types, however, a middle course can be steered which gives an impression of the extent of the unreliability of the diagnostic probability assessments for a new case with feature vector x. The approach follows the Bayesian device of using $p(\beta|D)$ as a vehicle for carrying the unreliability of the estimation process. The distribution $p(\beta|D)$ in its asymptotic Bayesian form is multivariate normal and so, for a given x, induces a multivariate normal distribution, say $N\{\mu(x), \sigma^2(x)\}$ on the score $\beta^T x$. This in turn induces, through the inverse logistic transformation (1·1), a logistic-normal distribution for the diagnostic probabilities $u_t = p(t|x,\beta)$.

For a case with given feature vector x the above argument leads to a predictive diagnostic probability for type I:

$$\int_{-\infty}^{\infty} \frac{e^v}{1+e^v} \phi\{v|\mu(x), \sigma^2(x)\}\, dv.$$

For any specified value, say 0·9, of this diagnostic probability there must be a relationship between corresponding $\mu(x)$ and $\sigma^2(x)$ and using Lauder's (1978, formula 3·7b) approximation this can be shown to be

$$\mu(x) = 2{\cdot}20\sqrt{\{1+0{\cdot}346\sigma^2(x)\}}.$$

It is thus possible for cases with the same predictive diagnostic probabilities to have widely different μ and σ^2 values and so different induced logistic-normal distributions for u_1. That such differences can reflect very different reliabilities of the diagnostic probability assertion is easily seen from Fig. 4 which shows the graphs of the logistic-normal distribution function

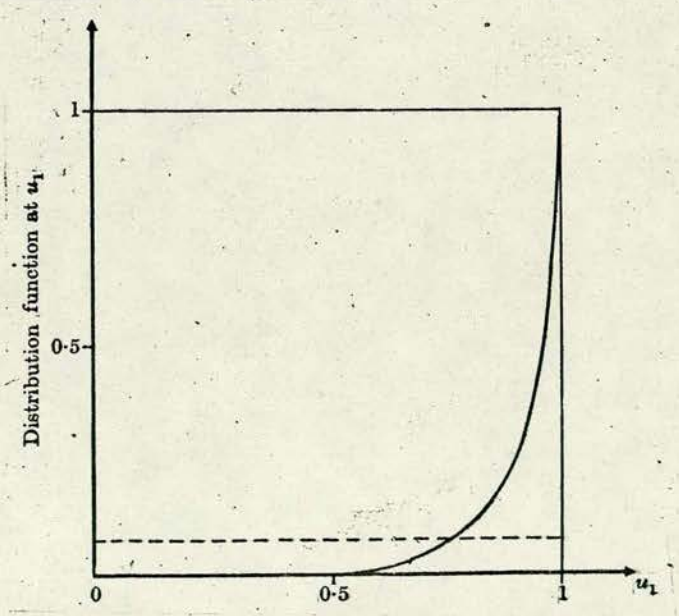

Fig. 4. Logistic normal distribution functions of the diagnostic probability u_1 for two new cases: case 1, ($\mu = 2{\cdot}9$, $\sigma = 1{\cdot}4$), shown by solid line; case 2, ($\mu = 105$, $\sigma = 81$), broken line.

of u_1 for two actual cases, with $\mu = 2{\cdot}9$, $\sigma = 1{\cdot}4$ and $\mu = 105$, $\sigma = 81$, each with predictive diagnostic probabilities of 0·9 for type 1. The first case gives a fairly reliable diagnosis for type I since there is little chance of u_1 being less than 0·5. The second case has obviously a very unreliable diagnosis since the logistic-normal probability distribution is almost entirely concentrated in the neighbourhoods of $u_1 = 0$ and $u_2 = 1$. Thus these induced logistic-normal distributions do give some insight into the diagnostic process.

5. Discussion

Although we have shown that the logistic-normal distributions provide a flexible tool for statistical analysis of a variety of applications a number of problems remain for future consideration.

(i) How can the techniques be adapted to cope with zero components in u vectors?

(ii) Can we develop satisfactory tests of the separate families, Dirichlet and logistic-normal, along the lines of Cox (1962)? In particular, to what extent are current tests of multivariate normality powerful against a Dirichlet alternative?

(iii) To what extent may logistic-normal distributions possess the component-additive property in some approximate form which would allow us to apply logistic-normal analysis to complete vectors of compositional data and to vectors collapsed through addition of components?

(iv) How worth while is it to widen the logistic-normal class, from the use of the logarithmic transformation to the complete Box & Cox (1964) class of transformations:

$$v_i = \{(u_i/u_{d+1})^\lambda - 1\}\lambda^{-1} \quad (i = 1, \ldots, d)?$$

We are currently investigating applications in such widely differing areas as petrology, soil compositions, fresh-water ecology and the analysis of subjective performance in inferential tasks. The answers to some of the above questions will clearly be conditioned by the particular needs of such practical problems.

References

Aitchison, J. (1974). Hippocratic inference. *Bull. Inst. Math. Applic.* **10**, 48–53.

Aitchison, J. (1976). Goodness of prediction fit. *Biometrika* **62**, 547–54.

Aitchison, J. & Begg, C. B. (1976). Statistical diagnosis when cases are not classified with certainty. *Biometrika* **63**, 1–12.

Aitchison, J. & Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

Aitchison, J., Habbema, J. D. F. & Kay, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statist.* **26**, 15–25.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Bloch, D. A. & Watson, G. S. (1967). A Bayesian study of the multinomial distribution. *Ann. Math. Statist.* **38**, 1423–35.

Box, G. E. P. & Cox, D. R. (1964). The analysis of transformations. *J.R. Statist. Soc.* B **26**, 211–52.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B **24**, 406–424.

De Finetti, B. (1972). *Probability, Induction and Statistics*. New York: Wiley.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* **36**, 149–76.

Kullback, S. & Liebler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 525–40.

Lauder, I. J. (1978). Computational problems in predictive diagnosis. *Compstat* 1978, 185–92.

Leonard, T. (1973). A Bayesian method for histograms. *Biometrika* **60**, 297–308.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35**, 1622–43.

Lindley, D. V., Tversky, A. & Brown, R. V. (1979). On the reconciliation of probability assessments (with discussion). *J. R. Statist. Soc.* A **142**, 146–80.

McCammon, R. B. (1975). *Concepts in Geostatistics*. New York: Wiley.

MOSIMANN, J. E. (1975). Statistical problems of size and shape. In *Statistical Distributions in Scientific Work*, Eds. G. P. Patil, S. Kotz and K. Ord, pp. 187–239. Dordrecht: Reidel.

TAYLOR, T. R., AITCHISON, J. & MCGIRR, E. M. (1971). Doctors as decision-makers: a computer-assisted study of diagnosis as a cognitive skill. *Br. Med. J.* **3**, 35–40.

WILKS, S. S. (1962). *Mathematical Statistics*. New York: Wiley.

[*Received March* 1979. *Revised January* 1980]

AITCHISON, J. (1980a)

A new approach to null correlations of proportions

Submitted to *J. Math. Geol.*

SUMMARY

Much work on the statistical analysis of compositional data has concentrated on the difficulty of interpreting correlations between proportions with an assortment of tests for null-correlations, for independence except for the constraint, F-independence of bounded variables, neutrality in the mean and in the median. This paper questions the appropriateness of characterising the dependence structure of proportions in terms of such concepts, suggests an alternative method of modelling, develops necessary distribution theory and tests, and illustrates the methodology in applications.

KEY WORDS: closed and open array, compositional data, correlations between proportions, Dirichlet distributions, logistic-normal distributions, petrogenesis, tests for basis independence.

INTRODUCTORY REVIEW OF THE PROBLEM

Some twenty years ago misinterpretations of correlations between proportions in the analysis of modal or compositional data became a matter of concern almost simultaneously in geology (Chayes, 1960, 1962; Krumbein, 1962; Chayes and Kruskal, 1966) and in biology (Mosimann, 1962, 1963). And still the debate continues (Butler, 1979). The difficulty arises because a basis or open vector of uncorrelated positive quantities $x_1,\ldots,x_{d+1}$ leads to a composition or closed vector of proportions $y_i = x_i/(x_1+\ldots+x_{d+1})$ $(i = 1,\ldots,d+1)$, which are necessarily correlated. How then are any apparent correlations in compositional data to be interpreted: as indicative of non-zero correlations in the basis or as merely induced through the process of forming a composition from an uncorrelated basis? The early work on this problem concentrated on determining the values of the induced or null correlations under a variety of model assumptions and on suggesting significance tests for comparing computed correlations against the null values. These tests were always presented with some hesitation for three important reasons.

(1) The distributions of the test statistics are not known (Mosimann, 1962, p.81; Chayes and Kruskal, 1966, p.696) and do not fall within the framework of any standard testing approach such as generalised likelihood ratio tests.

(2) The tests of null correlations are carried out separately for each pair of proportions. This procedure therefore is open to the same kind of criticism as the application of all $\frac{1}{2}k(k-1)$ t-tests of pairwise comparison of k treatments without a preliminary

overall F-test. The theory lacks the analogue of such an overall test (Chayes and Kruskal, 1966, p.696).

(3) When the tests detect non-null correlations it is by no means safe (Miesch, 1969) to conclude that the corresponding quantities in the basis are uncorrelated. Thus, despite the fact that the battery of pairwise tests, criticised in (2), is not designed as an overall test of the hypothesis that all correlations of the basis are zero this hypothesis is the only one which the battery effectively tests. No satisfactory analysis of the non-null case is available.

More recent work has largely been an attempt to introduce new concepts of non-association for proportions and relevant tests of significance: neutrality of one proportion with respect to another (Connor and Mosimann, 1969), neutrality in the mean (Darroch, 1969; Darroch and Ratcliff, 1970; Bartlett and Darroch, 1978), F-independence (Darroch and James, 1974), neutrality in the median (Darroch and Ratcliff, 1978). It remains to be seen whether these concepts, naturally more sophisticated than the concept of open correlations, prove straightforward enough for geologists to interpret. There is, however, a more fundamental difficulty. The properties of F-independence and neutrality lead almost inevitably to the description of variability through the Dirichlet class of distributions. The fact that, in the words of Darroch and James (1974, p.479), 'the Dirichlet distribution is almost the only one defined for continuous, positive, bounded-sum random variables which is easily handled for inference and descriptive purposes' then leads to the awkward question of how F-dependence and non-neutrality

are to be analysed.

This paper suggests that the unsatisfactory features of the above theories can be largely remedied by concentration on three fundamental aspects.

(a) A fuller appreciation of the relationship of closed to open variables or, in the terminology of this paper, of compositions to bases, and in particular the extent to which inferences can be made from compositional data to basis models.

(b) The introduction, as a consequence of (a), of a form of modelling which more simply, directly and tractably connects independence and non-association of the components of a basis to properties of the corresponding composition.

(c) The identification of a rich enough parametric class of distributions for compositional data which allows the description of both non-association and association within a single framework.

Since the measures of dependence used are covariances the main thrust of the paper may be seen as an attempt to resolve some of the difficulties of the null-correlation approach. As a bonus the resolution of (c) provides a tool which opens the way to further developments of the other approaches. For a more detailed discussion of some of these developments, see Aitchison (1980a).

THE RELATIONSHIPS OF BASES AND COMPOSITIONS

Algebraic considerations

From any (d+1)-dimensional vector x of positive quantities a d-dimensional vector y defined by

$$y_i = x_i/(x_1+\ldots+x_{d+1}) \quad (i = 1,\ldots,d)$$

can be formed. We then write y = C(x) and call y the composition of the basis x. We note that y is a vector of bounded sum 1 in the sense of Darroch and James (1974), since $e_d^T y \leq 1$, where e_d is the d-dimensional vector of units. Moreover, it is clear that y adequately describes the proportions of the constituents of x since the proportion of the (d+1)th constituent is $y_{d+1} = 1 - y_1 - \ldots - y_d$. The use of y rather than the augmented $\{y, y_{d+1}\}$ is mathematically more convenient since a composition is a d-dimensional rather than a (d+1)-dimensional entity, belonging to the d-dimensional simplex

$$S^d = \{y : y_i > 0 \quad (i = 1,\ldots,d), \ \sum_{i=1}^{d} y_i < 1\}.$$

It is well recognised that in many, probably most, geological applications an underlying basis is more conceptual than real, a convenient peg on which to hang discussion of non-association of proportions. What seems less clearly understood is the rather limited nature of the inferences possible about such conceptual bases from information on compositions. Starting at a purely algebraic level we see immediately that, given a composition y there is no way of uniquely reconstructing its basis. For if x is such a basis so that y = C(x) then, since C(zx) = C(x), where z

is a constant scalar or a random variable, we see that zx is also a basis. In fact the class of bases leading to a composition y is characterised by this multiplicative property, and the property of common composition defines equivalence classes of bases. In geometric terms for d = 1 the bases are points in the positive quadrant and compositions can be represented by points on the line segment from (1,0) to (0,1). The equivalence class of bases corresponding to a given composition y is formed by points on the ray from the origin through y.

Since a composition is uniquely, and most simply, specified by the values of $y_1,\ldots,y_d$ and hence $y_{d+1} = 1 - y_1 - \ldots - y_d$ the main thrust of the correlation approach to dependence has been in terms of correlations between y_i and y_j and the complications of interpretation. But the composition could equally well be specified in terms of any other d-dimensional vector v related to y through a one-to-one transformation. In searching for a sensible such transformation we should surely recognise the ability of a composition to determine a basis only up to a multiplicative factor. This immediately suggests the use of the ratio transformation $v_i = y_i/y_{d+1}$ $(i = 1,\ldots,d)$ or even, better, the logratio transformation

$$v_i = \log(y_i/y_{d+1}) = \log y_i - \log y_{d+1} \quad (i = 1,\ldots,d), \qquad (1)$$

since differences are usually simpler to handle than ratios. The inverse transformation of (1), from v to y, is the generalised logistic transformation

$$y_i = \exp(v_i)/\{1 + \sum_{i=1}^{d} \exp(v_i)\} \quad (i = 1,\ldots,d). \qquad (2)$$

This may seem at first sight an exotic tool for the analysis of compositions leading to very complicated interpretation problems. The opposite is the case, the transformation providing a natural and simple link between compositions and their equivalent bases and so probing to the root of the proportion correlation problem. The reason for such a simplification is not hard to find. The algebraic difficulty of compositions is their confinement to the simplex S^d, a difficult set to handle mathematically, whereas the corresponding space R^d of v vectors, the whole of d-dimensional real space, is a simpler space for analysis.

Expectation relationships

The relationship between the dependence structure of a basis x and its composition y takes a simple form when we work in terms of the equivalent v specification of the composition. In particular, if $u = \log x = \{\log x_1, \ldots, \log x_{d+1}\}$ then we can find very simple relationships between

$$\lambda = E(u), \quad \Omega = V(u)$$

and

$$\mu = E(v), \quad \Sigma = V(v).$$

Note that the dimensions of the vectors λ and μ are d+1 and d, and the dimensions of the matrices Ω and Σ are $(d+1) \times (d+1)$ and $d \times d$. Let A be the $d \times (d+1)$ matrix $[I_d - e_d]$ where I_d and e_d are the identity matrix and vector of units, each of dimension d. Then

$$\mu = A\lambda, \quad \Sigma = A\Omega A^T . \tag{3}$$

For a given μ and Σ we cannot find unique λ and Ω since, as we have seen earlier, there is an equivalence class of bases corresponding to a single composition. We can, however, easily identify the class of λ and Ω corresponding to a given μ and Σ. For the vector $\{y_1/y_{d+1},\ldots,y_d/y_{d+1},1\}$ forms a basis of y and so the general form of bases with composition y is $x = \{zy_1/y_{d+1},\ldots,zy_d/y_{d+1},z\}$, where z is any positive random variable. Since $\lambda = E(\log x)$ and $\Omega = V(\log x)$ we can see that the degree of arbitrariness is 1 for λ, represented by the arbitrary mean α of log z; and is d+1 for Ω, represented by the covariances $\beta_1,\ldots,\beta_d$ between log z and $\log(y_1/y_{d+1}),\ldots,\log(y_d/y_{d+1})$ and by the variance γ of log z. The appropriate expressions are then

$$\lambda = \begin{bmatrix} \mu + \alpha e_d \\ \alpha \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Sigma + e_d\beta^T + \beta e_d^T + \gamma U_d & \beta + \gamma e_d \\ \beta^T + \gamma e_d^T & \gamma \end{bmatrix},$$

where U_d is the $d \times d$ matrix of units.

Particular interest has been shown in the past in two related questions.

(1) What is the extent of the correlation or covariance structure induced by the constraining process of forming a composition from a basis? To what extent are correlations observed in compositions real or just induced by the constraint?

(2) How can we recognise from the covariance structure of a composition that it could have been produced from a basis of uncorrelated or independent components?

The strength of the present approach lies in the simplicity of the relationship for the circumstances described. For if a basis has independent components, then the logarithms of the components

are also independent and so $\Omega = \mathrm{diag}(\omega_1,\ldots,\omega_{d+1})$. Then the corresponding Σ_0 takes the form

$$\Sigma_0 = \mathrm{diag}(\omega_1,\ldots,\omega_d) + \omega_{d+1}U_d$$

$$= \begin{bmatrix} \omega_1+\omega_{d+1} & \omega_{d+1} & \cdots & \omega_{d+1} \\ \omega_{d+1} & \omega_2+\omega_{d+1} & \cdots & \omega_{d+1} \\ \omega_{d+1} & \omega_{d+1} & \ldots & \omega_d+\omega_{d+1} \end{bmatrix}. \qquad (4)$$

Since $\omega_1,\ldots,\omega_{d+1}$ are all positive we see that any composition corresponding to an independent basis must have equal positive covariances of the logratios v and that this common covariance must be less than every variance of a logratio v. We shall see later how we can construct from standard statistical theory a reasonable test of whether compositional data conform to this structure, that is whether the hypothesis H_0 of basis independence is tenable.

We reemphasise that even if we know that a composition has this special covariance structure all we can say is that the basis belongs to the equivalence class which contains an independent basis. Possible bases then have a covariance matrix of the special form

$$\begin{bmatrix} \omega_1+2\beta_1+\omega_{d+1} & \beta_1+\beta_2+\omega_{d+1} & \cdots & \beta_1+\beta_d+\omega_{d+1} & \beta_1+\omega_{d+1} \\ \beta_1+\beta_2+\omega_{d+1} & \omega_2+2\beta_2+\omega_{d+1} & \cdots & \beta_2+\beta_d+\omega_{d+1} & \beta_2+\omega_{d+1} \\ \beta_1+\beta_d+\omega_{d+1} & \beta_2+\beta_d+\omega_{d+1} & \cdots & \omega_d+2\beta_d+\omega_{d+1} & \beta_d+\omega_{d+1} \\ \beta_1+\omega_{d+1} & \beta_2+\omega_{d+1} & \cdots & \beta_d+\omega_{d+1} & \omega_{d+1} \end{bmatrix}.$$

Distributions for compositions

For a full analysis of compositional data some parametric class of distributions to describe the pattern of variability would be a clear advantage. Since compositions are elements or vectors in the simplex S^d the modelling problem is to find suitable distributions over this mathematically difficult space. The most, indeed the only, familiar class of distributions over S^d is the Dirichlet class with probability density functions

$$\frac{\Gamma(\alpha_1 + \ldots + \alpha_{d+1})}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_{d+1})} \prod_{i=1}^{d+1} y_i^{\alpha_i - 1},$$

where $\alpha = (\alpha_1, \ldots, \alpha_{d+1})$ is the parameter. The main difficulty with this family is its known mathematical property that such a compositional distribution can <u>always</u> arise from an <u>independent</u> basis, whose components are independent and gamma-distributed with equal 'scale' parameters.

Thus although the Dirichlet class can serve a useful purpose in providing a picture of the unavoidable and spurious correlations that arise simply out of the process of closure of an independent basis it cannot by its very nature be used to study proper covariance between components. For the description of real covariance a class of distributions over the simplex S^d with a much richer covariance structure is required. It is only recently (Aitchison & Shen, 1980) that such a class has been fully identified. A clue as to how to devise such a class and why the class turns out to be tractable in statistical analysis has been touched upon earlier. Since the awkward space S^d and the simple space R^d are related to each other

through the logistic and logratio transformations the following question immediately poses itself. If we start with a nice class of distributions in R^d, and what can be nicer than the multivariate normal class, and transform this to S^d through the logistic transformation, do we obtain a usable, tractable class in S^d, rich enough to describe distributions of real compositional data? We claim that the answer is certainly yes, and have set out the main interesting properties and some simple applications to compositional data in a previous paper (Aitchison & Shen, 1980). Since considerations here are directed towards presenting a new perspective on the problem of correlations in compositional data only the properties of this logistic-normal class of distributions which are of immediate interest for this particular problem are recorded.

If v in R^d is $N_d(\mu,\Sigma)$, that is d-dimensional multivariate normal with mean vector μ and covariance matrix Σ, then y in S^d, related to v through the logistic transformation (2), is said to follow a logistic-normal distribution, written $L_d(\mu,\Sigma)$ with parameters μ and Σ. Following a previous line of development we can show that the composition y of a basis x which is multivariate lognormally distributed, say $\Lambda_{d+1}(\lambda,\Omega)$ in the notation of Aitchison and Brown (1957), is $L_d(\mu,\Sigma)$ where μ and Σ are related to λ and Ω through (3). Indeed since multinormal, lognormal and logistic-normal are all defined in terms of these first and second-order moments (possibly of logarithms and logratios) all the comments on covariance structure of compositions and independence of bases carry through into this distributional form. In particular a logistic-normal composition $L_d(\mu,\Sigma)$ arising from a lognormal basis

with independent components must have Σ of the form (4). Thus we see that we have here the possibility of formulating a test of whether compositional data could be regarded as having arisen from independent bases. The question is translated into asking whether the covariance structure of the logratios of the proportions is consistent with, or is contrary to, the special structure (4) for Σ.

These two classes of distributions over S^d, the Dirichlet and the logistic-normal, are not unrelated. The logistic-normal is by far the richer and provides a stronger tool of statistical analysis and yet at the same time can be used as a substitute for any Dirichlet distribution. For Aitchison and Shen (1980) show that any Dirichlet distribution $D_d(\alpha)$ can be closely approximated by a logistic-normal distribution $L_d(\mu,\Sigma)$ where $\mu_i = \delta(\alpha_i) - \delta(\alpha_{d+1})$, $\sigma_{ii} = \varepsilon(\alpha_i) + \varepsilon(\alpha_{d+1})$, $\sigma_{ij} = \varepsilon(\alpha_{d+1})$ $(i \neq j)$ where δ and ε are the digamma function $\Gamma'(\cdot)/\Gamma(\cdot)$ and the trigamma function $\delta'(\cdot)$, respectively. Closeness is here judged in terms of the Kullback-Liebler (1951) measure of directed divergence of one density function from another. Not surprisingly the Σ for this closest logistic-normal distributes takes the basis-independence form (4). Thus we see that any statistical test of this particular covariance structure can be regarded as not only a test of the feasibility of an independent basis but also a test of whether the Dirichlet, the archetypal independence distribution, or the more general form of logistic-normal distribution is required.

Mosimann (1975b) has pointed out a property of lognormal bases which may have prevented previous consideration of the logistic-normal as a serious alternative to Dirichlet distributions. He shows that any lognormal basis x with a $\Lambda_{d+1}(\lambda,\Omega)$ distribution

with a composition having additive isometry (Mosimann, 1975b, p.223) or equivalently proportional invariance (Darroch and James, 1974, p.476), that is with a composition independent of $\sum_{i=1}^{d+1} x_i$, must be degenerate. At first sight it therefore appears that logistic-normal distributions should be applied only to situations where we are assured that the compositions need not satisfy such additive isometry; and as Darroch and James point out the use of compositional data often presupposes satisfaction of proportional invariance. It is perfectly possible, however, to have a non-degenerate logistic-normal distribution for a composition with the proportional invariance property satisfied, provided we do not insist on the basis itself being lognormal. There seem to be no strong grounds for insisting on such lognormality in the components. For example, it is easy to devise a petrogenetic model in which Σx_i turns out to be lognormal and the associated compositional distribution logistic-normal with additive isometry. More specifically the distribution with density function

$$\frac{\Sigma x_i}{(2\pi)^{\frac{1}{2}d} x_1 \ldots x_{d+1} |\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2} \sum_{i,j} \sigma^{ij} (\log \frac{x_i}{x_{d+1}} - \mu_i)(\log \frac{x_i}{x_{d+1}} - \mu_j) - \tfrac{1}{2}(\log \Sigma x_i - \alpha)^2/\omega^2\},$$

where $\Sigma^{-1} = [\sigma^{ij}]$, satisfies these requirements. There is no need even to insist on Σx_i having a lognormal distribution. The way therefore seems open for an investigation of the use of logistic-normal distributions in the analysis of compositional data. Aitchison and Shen (1980) have shown an application to discriminant analysis. Here we concentrate on the provision of a test of basis-indpendence.

AN OVERALL TEST OF BASIS INDEPENDENCE

The analysis of the previous section has led us inevitably to the problem of testing a null hypothesis H_0 that the covariance structure of Σ_0 is $\mathrm{diag}(\omega_1,\ldots,\omega_d) + \omega_{d+1} U_d$, where $\omega_i \geq 0$ $(i = 1,\ldots,d+1)$, the hypothesis of 'basis independence', against the alternative hypothesis that Σ takes a general positive definite form. The problem is similar to many considered in psychometric analysis, for example in Mukherjee (1970), except for the constraints on the positivity of the ω_i. The method adopted is the generalised likelihood ratio test, whose computation requires a simple iterative procedure. The simplicity depends on the special structure of the covariance matrix Σ_0 under the null hypothesis H_0. It is easy to show that

$$\det \Sigma_0 = \omega_1 \ldots \omega_{d+1} \left(\frac{1}{\omega_1} + \frac{1}{\omega_2} + \ldots + \frac{1}{\omega_{d+1}} \right),$$

$$\Sigma_0^{-1} = \mathrm{diag}(\tau_1,\ldots,\tau_d) - \left(\sum_{i=1}^{d+1} \tau_i \right)^{-1} \tau\tau^T,$$

where $\tau_i = \omega_i^{-1}$ $(i = 1,\ldots,d+1)$ and τ is the d-vector with components $\tau_1,\ldots,\tau_d$.

Suppose that the data D consist of n compositional vectors $y^{(1)},\ldots,y^{(n)}$ and that the matrix of corrected cross-products of logratio vectors $v^{(1)},\ldots,v^{(n)}$ is $V = [v_{ij}]$. The loglikelihood function, already maximised with respect to μ, can be expressed in the following way:

$$\log L = -\tfrac{1}{2} n \log |\Sigma| - \tfrac{1}{2} \mathrm{trace}(\Sigma^{-1} V).$$

Under the alternative hypothesis the maximising Σ is $\hat{\Sigma} = (1/n)V$. Under the null hypothesis H_0 the maximising process can easily be studied through the use of the standard Newton-Raphson iterative

methods, taking precautions to investigate the possibility that the maximising ω may be on one of the boundaries. The computational details are set out in the Appendix.

Let $\Lambda_0(D)$ denote the generalised likelihood ratio test statistic so obtained. Although there is little possibility of determining the exact distribution there is at least recourse to previous work on asymptotic distribution theory of generalised likelihood ratio test statistics. Since the constraints imposed by H_0 involve inequalities the standard results of Wald (1943) are not directly applicable, but require adjustment along the lines of Chernoff (1954) and Feder (1968). These adjustments are perhaps too complicated for general use and so we have chosen a simpler approach which allows us to wedge our problem between two other problems for which the standard theory applies.

Consider the two hypotheses

$$H_1 : \Sigma_1 = \text{diag}(\omega_1,\ldots,\omega_d) + \omega_{d+1} U_d,$$

without the restriction $\lambda_{d+1} \geqq 0$ on λ_{d+1}; and

$$H_2 : \Sigma_2 = \text{diag}(\omega_1,\ldots,\omega_d),$$

that is with $\lambda_{d+1} = 0$. It is clear that H_2 implies H_0 and that H_0 implies H_1 so that the corresponding generalised likelihood ratio test statistics satisfy

$$\Lambda_1(D) \leqq \Lambda_0(D) \leqq \Lambda_2(D)$$

for all data sets D. The hypothesis H_1 is a special form of a covariance structure studied by Mukherjee (1970) within the frame-

work of Wald (1943) theory. In the form here it places $\frac{1}{2}d(d-1) - 1$ constraint equations on the elements of Σ. Similarly H_2 falls trivially within general theory placing $\frac{1}{2}d(d-1)$ constraint equations on the elements of Σ. As already mentioned the appropriate asymptotic theory of Λ_0 under the null hypothesis is difficult, following the line of argument of Gleser and Olkin (1973) and involving mixtures of $\chi^2\{\frac{1}{2}d(d-1) - 1\}$ and $\chi^2\{\frac{1}{2}d(d-1)\}$ distributions. In order to avoid such complications and obtain a readily applicable test we appeal to the argument of embedding H_0 between H_1 and H_2 and use a playsafe critical value, namely the χ^2 value associated with $\frac{1}{2}d(d-1)$, the greater number of degrees of freedom.

Thus the asymptotic test we advocate computes $\Lambda_0(D)$ along the lines set out in the Appendix and then rejects the hypothesis H_0 of basis independence at significance level (at most α) when

$$\Lambda_0(D) > \chi^2\{\tfrac{1}{2}d(d-1);\alpha\},$$

where $\chi^2(\nu;\alpha)$ is the upper α point of the $\chi^2(\nu)$ distribution.

A simpler alternative to the above will often be sufficient for the practical purpose of rejecting basis independence. Since $H_0 \subset H_1$, rejection of H_1 implies rejecting of H_0, and so we may content ourselves by attempting to use the data D to reject the hypothesis H_1. The test statistic $\Lambda_1(D)$ is easily obtained by a modification of the $\Lambda_0(D)$ procedure in the Appendix, by removal of the insistence that $\omega_i \geq 0$ $(i = 1,\ldots,d+1)$.

SOME APPLICATIONS OF THE TEST

To illustrate the test of basis independence we use a number of applications which have appeared repeatedly in the literature to illustrate tests of null correlation. Because of the pairwise nature of the previous tests these have all tended to be low-dimensional; we therefore also undertake the analysis of some higher-dimensional data sets to emphasis the simplicity of the test procedure.

Fossil pollen counts. Mosimann (1962) has analysed data of Clisby and Sears (1955) giving 73 sets of the four proportions of fossil pollen grains of pine, fir, oak and alder. He tests for null correlations under a compound multinomial hypothesis and suspects that some of the correlations are significant but emphasises the rather tentative nature of his test procedures. In so far as the relation of oak and pine are concerned these data are also analysed through the concept of neutrality in the median by Darroch and Ratcliff (1978). One awkwardness of the logistic-normal analysis is that it cannot be applied to data with zero proportions. This difficulty has been circumvented by the admittedly ad hoc device of replacing zeroes by 0.005, and then readjusting the proportions to sum to unity. Application of the overall test of basis independence to this adjusted set of data with $n = 73$, $d = 3$ leads to comparison of the test quantity $\Lambda_0(D) = 11.01$ against critical $\chi^2(3)$ values of 7.81 at 5 per cent and of 11.34 at 1 per cent. Thus there is sufficient evidence at the 5 per cent significance level to reject the basis independence hypothesis associated with these compositional data.

If we concentrate on pine and oak, amalgamating fir and spruce, following the kind of neutrality investigations of Darroch and Ratcliff (1978), we find that the covariance matrix of log{pine/(fir + spruce)} and log{oak/(fir + spruce)} conforms to the basis independence pattern. Thus if interest is really in the composition of pine, oak and (fir + spruce) we cannot argue against the feasibility of basis independence. As we shall show elsewhere (Aitchison, 1980) this finding is not contrary to the finding of non-neutrality in the median by Darroch and Ratcliff (1978). The counterpart of their ideas within the logistic-normal framework is in the idea of conditional subcompositions, and basis independence does not in general imply independence of subcompositions.

Taupo volcanic rocks. Darroch and Ratcliff (1970, 1978) and Snow (1975) apply tests of neutrality in the mean and the median to two of the chemical components, SiO_2 and Al_2O_3, of 45 rock samples of the Taupo volcanic association reported by Steiner (1958). The nearest comparison with these analyses is an application of the test of basis independence to the 45 vectors of three proportions, SiO_2, Al_2O_3, remainder, so that $d = 2$. The test statistic $\Lambda_0(D) = 44.8$ is to be judged against $\chi^2(1)$ values. There is thus overwhelming evidence against the possibility of an underlying independent basis, a result in agreement with the earlier findings. But there is no need to confine ourselves to two of the components and the consequent lumping together of all the other components, an unnatural action if one raises the question of whether the lumping should be by weight or by volume. We have applied the test of H_0

to 44 of the vectors, omitting specimen no 10 because of its missing data, and lumping only the very minor minerals where none or only a trace was recorded. This procedure results in 13 components so that $d = 12$. Here $\Lambda_0(D) = 763$ to be tested against $\chi^2(66)$ values, so that again we have no hesitation in rejecting H_0.

Chemical variation in the Eocene lavas of the Isle of Skye. Thompson, Esson and Duncan (1972) present in their Table 2 chemical analyses, showing 10 components, of 32 basalts. Analysis of these for testing for the possibility of basis independence is readily carried out by the procedure of the Appendix and leads to a test quantity $\Lambda_0(D) = 329$, to be compared against upper $\chi^2(36)$ values. Again there is highly significant evidence against the basis independence hypothesis.

Sediment variability. McCammon (1975) sets a problem involving two data sets of (sand, silt, clay) compositions for (i) seven specimens of nearshore sediments, and (ii) ten specimens of offshore specimens. The first set yields a test quantity of $\Lambda_0(D) = 0$, seen to be absolutely reasonable when we note that the covariance matrix of the two logratios $v_1 = \log(x_1/x_3)$, $v_2 = \log(x_2/x_3)$ is

$$\begin{bmatrix} 0.731 & 0.216 \\ 0.216 & 0.494 \end{bmatrix}$$

readily conforming to the pattern

$$\begin{bmatrix} \omega_1 + \omega_3 & \omega_3 \\ \omega_3 & \omega_2 + \omega_3 \end{bmatrix},$$

with ω_1, ω_2, ω_3 all positive.

For the offshore specimens the covariance matrix

$$\begin{bmatrix} 2.978 & 0.754 \\ 0.754 & 0.453 \end{bmatrix}$$

does not conform to the above pattern, but still the test quantity $\Lambda_0(D) = 1.10$ is not sufficiently large to allow rejection of H_0. Thus for both nearshore and offshore data we cannot refute the hypothesis that the composition has its origin in a basis with independent components.

DISCUSSION

The test of basis independence developed and applied in this paper has several advantages over previous attempts at analysis. It is based on a simple and natural way of linking the dependence structures of open and closed models, it provides an overall test of the complete structure as opposed to separate pairwise tests, its critical value is relatively well based in asymptotic test theory compared with the tentative nature of many previous tests. More important, however, is that the development does not stop at the test of basis independence but allows further investigation in the event of rejection of the hypothesis. There are many possibilities of further investigation through logistic-normal distributions. There may be other patterned covariance structures, depicting some special forms of dependence, which could be next investigated, along lines similar to those investigated by psychologists; see, for example, Mukherjee (1970) and Gleser and Olkin (1973). Interest may be directed towards subcompositions of the whole vector in the sense of Aitchison and Shen (1980), and the investigation of whether the subcompositions could have independent bases. Alternatively we may be more concerned with some form of statistical principal component analysis and investigation of the effective dimensionality of the pattern of variability. All of this is in striking contrast to other formulations which fail to provide any statistical framework for the quantitative investigation of truly correlated proportions.

We have concentrated on one aspect of the value of the new approach, a clearer understanding of the nature of null correlations

and a practical tool for their analysis. There are many other clarifications and methodology which emerge immediately. For example, the relation of basis independence to other forms of non-association throws some interesting light on modelling. We can actually devise a simple test for proportional invariance, size homogeneity, or additive isometry (Aitchison, 1980b). Finally since the logistic-normal has a limit law similar to the central limit theorem for normal distributions, interesting questions can be raised about the possibility of providing a genetic explanation, along the lines of the lognormal genesis by breakage for particle size distributions, of the occurrence of logistic-normal compositional patterns.

APPENDIX: THE COMPUTATION OF THE TEST STATISTIC

Let D denote the data set consisting of n compositional vectors $x_1,\ldots,x_n$ in S^d, with corresponding logratio vectors $v_1,\ldots,v_n$. Let $\bar{v}$ denote the mean vector and $V = \sum_{i=1}^{n} (v_i - \bar{v})(v_i - \bar{v})^T$ the matrix of cross products. As pointed out in the main test, in §3, the likelihood under the logistic-normal model is maximised at $\hat{\mu} = \bar{v}$, $\hat{\Sigma} = V/n$, and the only awkward problem is the maximisation under the basis independence hypothesis that the covariance structure is of the form:

$$\sigma_{ii} = \omega_i + \omega_{d+1} \qquad (i = 1,\ldots,d),$$

$$\sigma_{ij} = \omega_{d+1} \qquad (i \neq j),$$

with $\omega_i \geqq 0$ $(i = 1,\ldots,d+1)$.

As far as investigating the region $\omega_i \geqq 0$ $(i = 1,\ldots,d+1)$ is concerned we can show, after some tedious algebra, that the $(d+1)$-vector $d(\lambda)$ of derivatives of the loglikelihood has ith component

$$d_i(\omega) = \tfrac{1}{2}n(\omega_i - \tau) - \tfrac{1}{2}v_{ii} + \tau \sum_{j=1}^{d+1} v_{ij}/\omega_j - \tfrac{1}{2}\tau^2 \sum_{i=1}^{d+1}\sum_{j=1}^{d+1} v_{ij}/(\omega_i\omega_j),$$

where $\tau^{-1} = \omega_1^{-1} + \ldots + \omega_{d+1}^{-1}$, and that the $(d+1) \times (d+1)$ information matrix is

$$B(\omega) = \tfrac{1}{2}n\{\mathrm{diag}(\omega_1^2 - 2\tau\omega_1,\ldots,\omega_d^2 - 2\tau\omega_d) + \tau^2 U_d\}.$$

The usual iterative procedure leading from rth iterate $\omega^{(r)}$ $(r = 0,1,\ldots)$ to the $(r+1)$th iterate $\omega^{(r+1)}$ is:

$$\omega^{(r+1)} = \omega^{(r)} + [B(\omega^{(r)})]^{-1}\, d(\omega^{(r)}).$$

Recommended initial values are

$$\omega_{d+1}^{(0)} = \sum_{i=1}^{d}\sum_{j=i+1}^{d} v_{ij}/\{\tfrac{1}{2}nd(d-1)\}, \omega_i^{(0)} = v_{ii}/n - \omega_{d+1}^{(0)} \quad (i = 1,\ldots,d) \quad (5)$$

if these are all non-negative. Otherwise, if the above $\omega_{d+1}^{(0)} < 0$ set

$$\omega_{d+1}^{(0)} = 0.001, \ \omega_i^{(0)} = v_{ii}/n \quad (i = 1,\ldots,d)$$

and if $\omega_{d+1}^{(0)}$ in (5) is non-negative and $\omega_j^{(0)}$ is the minimum negative value of $\omega_i^{(0)}$ $(i = 1,\ldots,d)$, set

$$\omega_{d+1}^{(0)} = v_{jj}/n, \ \omega_j^{(0)} = 0.001, \ \omega_i^{(0)} = (v_{ii} - 2v_{ij} + v_{jj})/n. \ (i \neq j, d+1)$$

As each iterative stage is completed it is easy to check whether all $\omega_i^{(r+1)}$ are positive. If not set, any which are negative to 0.001 before the next iterative cycle. If the maximisation is on the boundary, that is with some ω_i $(i = 1,\ldots,d+1)$ zero then the above procedure picks up this fact by the corresponding iterate becoming smaller and smaller. As a check on this, any case where some ω_i is zero is very simply solved, with for $\omega_{d+1} = 0$,

$$\hat{\omega}_i = v_{ii}/n$$

and for ω_i $(i \neq d+1) = 0$

$$\hat{\omega}_{d+1} = v_{ii}/n, \ \omega_j = (v_{ii} - 2v_{ij} + v_{jj})/n.$$

A simple program for the test procedure has been written in BASIC and implemented on the Wang 2200S minicomputer system. A listing is available from the author on request.

REFERENCES

Aitchison, J., 1980, Distributions on the simplex: Paper to be presented at the International Summer School on Statistical Distributions in Scientific Work, Trieste July 1980.

Aitchison, J., 1980, Testing for additive isometry and proportional invariance: submitted to Biometrics.

Aitchison, J., and Brown, J.A.C., 1957, The Lognormal Distribution: Cambridge University Press.

Aitchison, J. and Shen, S.M., 1980, Logistic-normal distributions: some properties and uses: Biometrika 67, to appear.

Bartlett, N.R. and Darroch, J.N., 1978, Regression and correlation of bounded-sum variables: Vistelius Commemoration Volume.

Butler, J.C., 1979, Trends in ternary petrologic variation diagrams - fact or fantasy?: Amer. Mineralogist, v.64, p.1115-1121.

Chayes, F., 1960, On correlation between variables of constant sum: Jour. Geophys. Res., v.65, p.4185-4193.

Chayes, F., 1962, Numerical correlation and petrographic variation: Jour. Geol., v.70, p.440-452.

Chayes, F. and Kruskal, W., 1966, An approximate statistical test for correlations between proportions: Jour. Geol., v.74, p.692-702.

Chernoff, H., 1954, On the distribution of the likelihood ratio: Ann. Math. Stat., v.25, p.573-578.

Clisby, K.H., and Sears, P.B., 1955, Palynology in southern North America. Part III: Microfossil profiles under Mexico City correlated with sedimentary profiles: Geol. Soc. America Bull., v.66, no. 5, p.511-520.

Connor, J.R., and Mosimann, J.E., 1969, Concepts of independence for proportions with a generalization of the Dirichlet distribution: Jour. Amer. Statist. Assoc., v.64, p.194-206.

Darroch, J.N., 1969, Null correlations for proportions: Jour. Math. Geol., v.1, no. 2, p.221-227.

Darroch, J.N., and James, I.R., 1974, F-independence and null correlations of continuous, bounded-sum, positive variables: Jour. Roy. Stat. Soc., Ser. B, v.36, no. 3, p.467-483.

Darroch, J.N., and Ratcliff, D., 1970, Null correlations for proportions II: Jour. Math. Geol., v.2, p.307-312.

Darroch, J.N., and Ratcliff, D., 1971, A characterization of the Dirichlet distribution: Jour. Amer. Stat. Assoc., v.66, p.641-643

Darroch, J.N., and Ratcliff, D., 1978, No-association of proportions: Jour. Math. Geol., v.10, p.361-368.

Feder, P.I., 1968, On the distribution of the loglikelihood ratio test statistic when the true parameter is 'near' the boundaries of the hypothesis regions: Ann. Math. Stat., v.39, p.2044-2055.

Gleser, L.J., and Olkin, I., 1973, Multivariate statistical inference under marginal structure. Brit: Jour. Math. Stat. Psychol., v.26, p.98-123.

Krumbein, W.C., 1962, Open and closed number systems stratigraphic mapping: Bull. Amer. Assoc. Petrol. Geologists, v.46, p.2229-2245.

Kullback, S., and Liebler, R.A., 1951, On information and sufficiency: Ann. Math. Stat., v.22, p.525-540.

Miesch, A.T., 1969, The constant sum problem in geochemistry: In Merriam, D.F. (ed.), Computer Applications in the Earth Sciences, New York: Plenum Press, p.161-177.

Mosimann, J.E., 1962, On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions: Biometrika, v.49, p.65-82.

Mosimann, J.E., 1963, On the compound negative multinomial distribution and correlations among inversely sampled pollen counts: Biometrika, v.50, p.47-54.

Mosimann, J.E., 1975a, Statistical problems of size and shape. I. Biological applications and basic theorems, in Patil, G.P., Kotz, S., and Ord, J.K., (eds.), Statistical distributions in scientific work, v.2: D. Reidel Publishing Company, Dordrecht, Holland, p.187-217.

Mosimann, J.E., 1975b, Statistical problems of size and shape. II. Characterizations of the lognormal, gamma and Dirichlet distributions, in Patil, G.P., Kotz, S., and Ord. J.K. (eds.), Statistical distributions in scientific work, v.2: D. Reidel Publishing Company, Dordrecht, Holland, p.219-239.

Mukherjee, B.N., 1970, Likelihood ratio tests on statistical hypotheses associated with patterned covariance matrices in psychology: Brit. Jour. Math. Stat. Psychol., v.23, p.89-120.

Snow, J.W., 1975, Association of proportions: Jour. Math. Geol., v.7, no. 1, p.63-73.

Steiner, A., 1958, Petrographic implications of the 1954 Ngauruhoe lava and its xenoliths: New Zealand Jour. Geol. Geophys., v. 1, p.325-363.

Thompson, R.N., Esson, J., and Duncan, A.C., 1972, Major element chemical variation in the Eocene lavas of the Isle of Skye, Scotland: Jour. Petrology, v.13, p.219-253.

Wald, A., 1943, Tests of statistical hypotheses concerning several parameters when the number of observations is large: Trans. Amer. Math. Soc., v.54, p.426-482.

AITCHISON, J. (1980b)

Testing for additive isometry and proportional invariance

Submitted to *Biometrics*

Summary

A reevaluation of the property of additive isometry in size and shape studies, and of proportional invariance in the analysis of compositional data is undertaken. Particular emphasis is placed on correcting a generally held view that the lognormal model can only sustain the hypothesis of additive isometry or proportional invariance at the expense of complete degeneracy. A new non-degenerate lognormal, or more correctly logistic-normal, model is formulated which not only allows additive isometry and proportional invariance but also provides a simple test of these hypotheses within the model. The model and the test are illustrated on two biological data sets.

1. Introduction

In the study of size and shape associated with a set of d+1 measurements $y_1,\ldots,y_{d+1}$ shape is usually defined as the measurement-dimensionless vector $x = (x_1,\ldots,x_d)$, where

$$x_i = y_i / \sum_{j=1}^{d+1} y_i \quad (i = 1,\ldots,d) \tag{1}$$

or some simple one-to-one transformation of it such as

$$w_i = x_i/(1 - x_1 - \ldots - x_d) \quad (i = 1,\ldots,d). \tag{2}$$

When size is regarded as $z = y_1 + \ldots + y_{d+1}$ then additive isometry (Mosimann 1975a) is defined as the statistical independence of shape x and size z. Sprent (1972) presents a survey of such ideas and relates them to their historical development, while Mosimann (1970, 1975a, 1975b) considers the mathematical modelling of this and related aspects.

A conceptually similar problem occurs in another common area of application, the analysis of compositional data, where some total such as the total urinary excretion of steroid metabolites in 24 hours, say z, is broken down into the quantities excreted of d+1 component steroid metabolites $y_1,\ldots,y_{d+1}$. Here the total excretion z is the obvious counterpart of size while the proportional

---

Key Words: Additive isometry; Compositional data; Logistic-normal distribution; Lognormal distribution; Proportional invariance; Size and shape.

composition $x_i = y_i/(y_1 + \ldots + y_{d+1})$ $(i = 1,\ldots,d)$ corresponds to the shape vector. Again interest may focus on whether composition is independent of size. This question of proportional invariance (Darroch and James, 1974) is often very important, taking practical expression in such questions as the following. If we concentrate our analysis on the variability of composition are we in fact losing information by neglecting size? Can we discriminate between types of disease through the use of compositional proportions of steroid metabolites, or should we also take into account the absolute magnitude of the excretions?

One of the major hurdles to the statistical analysis of size and shape, and equally to compositional data, has been the absence of a versatile enough class of distributions to describe the pattern of variability of shape x, as defined in (1), or w as defined in (2), and its relationship to additive size z. In the x form we are concerned with distributions over the d-dimensional simplex

$$S^d = \{x : x_i > 0 \quad (i = 1,\ldots,d), \sum_{i=1}^{d} x_i < 1\}, \tag{3}$$

and the popular Dirichlet class is unrealistic in that it is essentially associated with <u>independent</u> Gamma-distributed measurements $y_1,\ldots,y_{d+1}$. Darroch and James (1974) regret the fact that there seems to be a scarcity of distributions capable of describing situations which have any real structure of association. In the w form the multivariate lognormal model has been considered particularly by Mosimann (1975a, 1975b) but, as far as additive isometry is concerned, dismissed because of a degeneracy property that additive isometry implies. This is unfortunate since a very minor

modification to the modelling removes this defect and provides a rich class of distributions, closely related to the lognormal class, to describe the joint distribution of x, or w, and z. Whether or not x is additively isometric with respect to z depends on some parameters of the distribution, and hence the way is open for the development of a test of the hypothesis of additive isometry. In the event of the distribution having the additive isometry property there is no necessity for it to be degenerate.

In §2 we present the model explaining carefully how it avoids the force of Mosimann's theorem on degeneracy. In §3 we devise a test for additive isometry, in §4 show how it applies in two biological situations, and in §5 indicate a development of this form of analysis.

2. Logistic-Normal and Lognormal Models

A central result on lognormal models is that reported by Mosimann (1975b) as his Theorem 1: if $y = (y_1,\ldots,y_{d+1})$ is multivariate lognormal and the d-vector x defined by (1) is independent of $z = y_1 + \ldots + y_{d+1}$ then the distribution of y is degenerate, the covariance matrix of log y being proportional to a $(d+1) \times (d+1)$ matrix with all elements equal to 1. As indicated in §1 the discovery of this result has apparently removed lognormal distributions from modelling considerations when additive isometry or proportional invariance is required.

Let us, however, reconsider the essential features of the lognormal model. If y is (d+1)-dimensional lognormal then x follows a logistic-normal distribution $L_d(\mu,\Sigma)$ on the simplex S^d, as defined by Aitchison and Shen (1980), with density function of the form

$$p(x) = (2\pi)^{-\frac{1}{2}d}|\Sigma|^{-\frac{1}{2}}(x_1\ldots x_{d+1})^{-1}$$
$$\exp[-\tfrac{1}{2}\{\ell n(x/x_{d+1})-\mu\}^T\Sigma^{-1}\{\ell n(x/x_{d+1})-\mu\}], \qquad (4)$$

where $\ell n(x/x_{d+1})$ is the d-dimensional vector with ith component $\ell n(x_i/x_{d+1})$. The change in modelling strategy now advocated is to take this shape or compositional distribution p(x) as starting position. Then, if q(z) is <u>any</u> density function over the positive real numbers, the joint distribution of x and z with density function

$$p(x)q(z) \qquad (5)$$

is non-degenerate and obviously possesses the property of additive isometry. This has been achieved simply by not insisting that $y_i = zx_i$

$(i = 1,\ldots,d+1)$ are lognormal, and, of course, there is no fundamental reason why in practical modelling we should insist on y being multivariate lognormal. Historically the lognormal distribution has been seen as a convenient starting point in modelling because of its familiarity, but, now that the logistic-normal class on S^d is defined and available to describe shape or composition there is no theoretical reason why it should not form the basis of modelling. With the model (5) we can, by using the transformation $y_i = zx_i$ $(i = 1,\ldots,d+1)$ with Jacobian $Dy/D(x,t) = z^d = (\Sigma y_i)^d$, arrive at the corresponding density function

$$(\Sigma y_i)^{-d} p(y/\Sigma y_i) q(\Sigma y_i) \tag{6}$$

for y.

We can even bring the model closer to Mosimann's lognormal model by supposing that the size $z = \Sigma y_i$ follows a lognormal distribution, say $\Lambda_1(\gamma,\delta^2)$, in the notation of Aitchison and Brown (1957). The density function (6) for y then takes the form

$$(2\pi)^{-\frac{1}{2}(d+1)} \delta^{-1} |\Sigma|^{-\frac{1}{2}} \Sigma y_i (y_1 \ldots y_d y_{d+1})^{-1}$$

$$\times \exp[-\tfrac{1}{2} \sum_{i,j=1}^{d} \sigma^{ij} \{\log(y_i/y_{d+1}) - \mu\}\{\log(y_i/y_{d+1}) - \mu\} - \tfrac{1}{2}(\log \Sigma y_i - \gamma)^2/\delta^2], \tag{7}$$

where σ^{ij} is the (i,j)th element of Σ^{-1}. This seems as attractive a modelling assumption as the multivariate lognormal distribution assumption. In biological terms we are first saying that the overall size of the organism or object of study varies according to a lognormal pattern, and there are many explanations - law of proportionate effect, theory of breakage - in support of such a possibility. Secondly we are saying that the internal structure, the division

of total size into its d+1 components, follows a logistic-normal distribution. It is possible to devise a random mechanism, similar to the central limit property, which generates logistic-normal distributions (Aitchison, 1980). Briefly this envisages a process with internal composition perturbed in a series of stages, with the (r+1)th stage components $x_i^{(r+1)}$ arising from the rth stage components $x_i^{(r)}$ through a relationship

$$x_i^{(r+1)} \propto z_i^{(r)} x_i^{(r)} \tag{8}$$

where the random vector $z^{(r)}$ of perturbations does not necessarily consist of independent components. There is, of course, no way of verifying such a formative action, but this is equally a problem with the multivariate lognormal model. All we are claiming here is that there could be plausible reasons for investigation of this new form of model.

Our modelling so far has provided only for additive isometry and proportional invariance within a non-degenerate model. There is, however, no reason why at (5) we should not adopt as a starting point a pattern of variability for x conditional on z, with conditional density function $p(x|z)$, so that the joint distribution of x and z is

$$p(x|z)q(z). \tag{9}$$

It is then natural to continue to take $p(x|z)$ of logistic-normal form, with mean dependent on z, say $L_d(\alpha+\beta z,\Sigma)$. The case $\beta = 0$ corresponds to additive isometry and proportional invariance and so the testing of the hypothesis that $\beta = 0$ forms an extremely con-

venient method of testing the hypotheses of additive isometry and proportional invariance.

3. A Statistical Test for Additive Isometry and Proportional Invariance

We suppose that we have a data set D consisting of n shape or compositional vectors $x_1,\ldots,x_n$ and corresponding (additive) sizes $z_1,\ldots,z_n$. The problem is then how to use this data set D to test the hypothesis that the d-vector $\beta = 0$ on the basis of a logistic-normal model $L_d(\alpha+\beta z,\Sigma)$ for the conditional distribution of x for given z. The great advantage of the logistic-normal distribution is its tractability in statistical analysis since a logratio transformation $v_{ij} = \log\{x_{ij}/(1 - x_{i1} - \ldots - x_{id})\}$ $(j = 1,\ldots,d)$ transforms each x_i to a corresponding v_i which is $N_d(\alpha+\beta z_i,\Sigma)$. In terms of the transformed data $v_1,\ldots,v_n$ we are faced with standard testing of a linear hypothesis within the context of analysis of dispersion in multivariate normal theory (Rao, 1965, Chapter 8). Write $\bar{v} = \Sigma v_i/n$, the matrix of cross-products $R = \Sigma(v_i - \bar{v})(v_i - \bar{v})^T$, and the maximum likelihood estimate $\hat{\beta} = \Sigma(z_i - \bar{z})(v_i - \bar{v})/\Sigma(z_i - \bar{z})^2$ of β. Then standard test theory (Rao, 1965, Table 8c.5β) leads to an exact test of $\beta = 0$ by comparison of

$$\frac{|R| - |R - \hat{\beta}\hat{\beta}^T\Sigma(z_i - \bar{z})^2|}{|R - \hat{\beta}\hat{\beta}^T\Sigma_1(z_i - \bar{z})^2|} \frac{n - d - 1}{d} \qquad (10)$$

against critical values of the F distribution at d and n-d-1 degrees of freedom.

To demonstrate the simplicity of the test we present in the next section two applications, the first to the investigation of additive isometry in a shape and size analysis and the second to the question of proportional invariance in an analysis of compositional data.

4. Applications

4.1. Radiological Heart Measurements

The ratio of 'heart width' to 'thorax width' as seen on a heart X-ray is considered by some radiologists to have appreciable diagnostic value in detecting certain types of hypertensive conditions, and attempts have been made to quantify the ratio by radiologists actually making measurements on heart X-rays as displayed on a viewing screen. In any consideration of the diagnostic value of such a ratio it is obviously of some interest to know whether or not its distribution is independent of size of person, normal or hypertensive. If there is such isometry then the use of the ratio is well founded, provided of course that its distributions for normal and hypertensive persons have a reasonable degree of separation. If there is no such isometry then it may well be more appropriate for diagnostic purposes to take both measurements into consideration rather than reduce them to the ratio.

Fig. 1 shows the scattergram for 75 heart X-rays, with the measurements y_1 = heart width and y_2 = thorax width - heart width, so that $z = x_1 + x_2$ = thorax width is being interpreted as size. Of the 75 patients 20 were normal and 55 were hypertensive. The diagnostic ratio $x = y_1/(y_1 + y_2)$ is constant along rays through the origin. Tests of additive isometry as described in §3 have been carried out separately on the normal and hypertensive groups and also on the combined group of 75 patients. Here the three models involved are respectively that x is $L_1(\alpha_1+\beta_1 z,\sigma_1^2)$, $L_1(\alpha_2+\beta_2 z,\sigma_2^2)$, $L_1(\alpha+\beta z,\sigma^2)$. The test quantities (10) for these three groups are 7.85, 4.96, 9.49 to be compared against F(1,18), F(1,53), F(1,73)

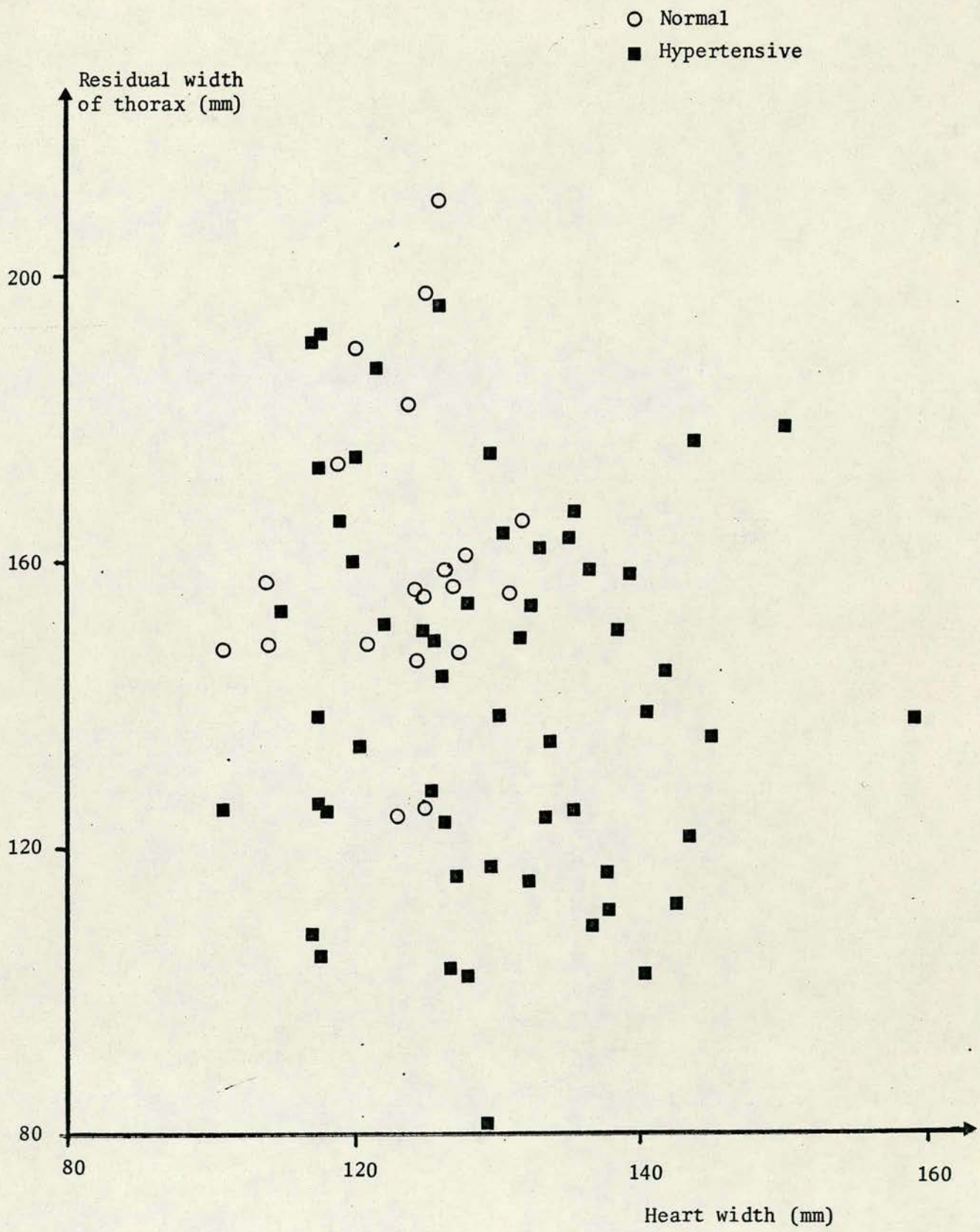

Figure 1

Two measurements of heart X-rays for 20 normal
and 55 hypertensive persons

critical values, and so there is significant departure from the hypothesis of additive isometry at the 5,5 and 1 per cent significance levels.

Also of some interest is whether the lack of additive isometry is identical in the normal and hypertensive groups: in other words, is the hypothesis that $\alpha_1 = \alpha_2(=\alpha)$ and $\beta_1 = \beta_2(=\beta)$ tenable. A test of this is equivalent to a standard normal linear regression test of the equality of two regression lines, since $\log\{x/(1-x)\}$ is $N(\alpha_i + \beta_i z, \sigma_i^2)$ $(i = 1,2)$. The appropriate test quantity is 5.35 to be compared against F(2,71) critical values and so the forms of dependence on additive size are significantly different at the 0.1 per cent level. This difference suggests that the diagnostic ratio has some diagnostic merit, but the absence of additive isometry further suggests that there is a possibility of an improved diagnostic system if we take account of both measurements y_1 and y_2 rather than their ratio. Since our purpose is the limited one of illustrating a simple test of additive isometry within the lognormal and logistic-normal framework we shall not pursue this possibility further here.

4.2. Urinary Steroid Metabolite Excretions

As part of a study into the differential diagnosis of Cushing's syndrome the amounts of 14 steroid metabolites excreted in urine over 24 hours have been recorded by Nielsen (1971) in Cushing syndrome patients and in 37 normal adults and in 30 normal children. A successful differential diagnostic system has been built up on a training set based on adult amounts but this is obviously not applicable to children since their excreted amounts are on average appreciably less than the corresponding amounts for adults. Direct

application to 'outlier' children is unsound statistical practice. For example, one two-year old child processed through the adult diagnostic system is placed with almost certainty into the adenoma (benign tumour) class when in fact the correct diagnosis is a carcinoma of the adrenal gland. A question which therefore arises in such applications is whether absolute amounts are really necessary or whether proportions would be adequate, that is whether the compositions satisfy proportional invariance. To undertake a discussion of this in the full differential diagnostic context would take us well beyond the scope of this present paper and so we confine ourselves to the examination of the problem of proportional invariance in relation to the set of normal adults and normal children. Moreover, in order to be able to present a complete picture of our tests here we condense the data for each case into a three-dimensional vector with three natural groupings of metabolites, (i) total cortisol metabolites, (ii) total corticosterone metabolites, (iii) remainder (pregnanetriol + Δ-5-pregnentriol). The data are given in Table 1.

The test applied to normal adults produces a test quantity 1.35 to be compared against F(2,34) critical values; also for normal children the test quantity is 1.85 to be compared against F(2,27) critical values. In neither case can we reject the hypothesis of proportional invariance. This is perhaps to be expected since within each of these groups there is not large variability in size, certainly not in relation to the differences between the groups. Moreover, the motivation for the proportional invariance study is to see whether this property extends to the combined group of adults and children. For this combined group of 67 persons the test

TABLE 1

Urinary Excretions (mg/24 hr) of Steroid Metabolites for Normal Adults and Normal Children

Adults			Children		
(i)	(ii)	(iii)	(i)	(ii)	(iii)
2.47	0.29	0.40	1.78	0.29	0.075
2.96	0.39	1.10	1.77	0.21	0.065
4.09	0.26	0.90	2.01	0.37	0.045
3.27	0.24	1.80	1.17	0.25	0.025
2.30	0.51	0.50	1.29	0.17	0.055
5.06	0.50	1.30	2.80	0.26	0.305
2.86	0.42	1.50	1.36	0.30	0.205
3.38	0.51	0.60	3.31	0.28	0.205
3.18	0.12	0.50	0.52	0.07	0.005
3.49	0.31	1.30	1.97	0.14	0.005
3.56	0.57	1.20	2.10	0.23	0.125
5.24	0.51	0.48	2.41	0.36	0.055
3.62	0.29	0.50	1.60	0.26	0.065
2.99	0.38	0.50	1.14	0.066	0.015
2.18	0.34	0.60	1.44	0.163	0.075
3.86	0.37	0.60	1.96	0.21	0.145
3.04	0.35	0.40	2.01	0.27	0.105
2.82	0.29	0.60	0.83	0.12	0.045
2.40	0.38	0.90	1.58	0.12	0.105
4.73	0.35	1.40	1.82	0.20	0.105
3.49	0.40	0.80	1.49	0.25	0.045
6.32	0.86	2.30	1.96	0.23	0.115
3.88	0.37	0.90	1.97	0.29	0.105
3.79	0.42	1.20	0.831	0.10	0.105
9.95	1.00	0.80	1.58	0.08	0.215
7.03	0.56	1.10	2.84	0.09	0.105
4.23	0.48	1.20	1.77	0.14	0.085
5.60	0.48	1.80	3.02	0.34	0.505
4.30	0.36	0.60	1.17	0.17	0.205
9.74	0.76	1.10	1.69	0.27	0.085
4.54	0.29	0.60			
6.33	0.92	0.90			
6.65	0.66	2.00			
5.96	0.50	1.90			
1.86	0.50	0.50			
1.33	0.13	0.20			
5.42	0.46	0.60			

(i) Total cortisol metabolites

(ii) Total corticosterone metabolites

(iii) Pregnanetriol + Δ-5-pregnentriol

quantity is 15.1 to be compared against F(2,64) critical values. This is highly significant ($P < 0.001$) and so the hypothesis of proportional invariance for this overall set of data must be firmly rejected. As we have said earlier it is not our purpose here to discuss the consequences of this for the differential diagnosis problem.

We should perhaps also make the comment that when the full 14-dimensional compositional structure is analysed it is found that the hypothesis of proportional invariance is rejected for each of the groups normal adults and normal children separately, each at the 1 per cent significance level. Investigation of the combined group reinforces this conclusion, rejecting the hypothesis of proportional invariance at the 0.01 per cent level.

5. Discussion

We have shown how the advent of the class of logistic-normal distribution opens the way to an extremely simple test of additive isometry and proportional invariance, achieved through the simple relationship of the logistic-normal to its associated multivariate normal counterpart. This relationship can indeed be further exploited if there is any possibility that shape or composition may be dependent on some concomitant variable or vector u. Within a general model which specifies the distribution of shape or composition x, conditional on size z and concomitant information u, as $L_d(\alpha+\beta z+\Gamma u,\Sigma)$ we may proceed to test a lattice of hypotheses such as

(i) $\Gamma = 0$, the hypothesis of no dependence on the concomitant information;

(ii) $\beta = 0$, the hypothesis of conditional additive isometry, that is, for given u, shape or composition x is independent of additive size z;

(iii) $\beta = 0$, $\Gamma = 0$, the hypothesis of complete independence of shape or composition on additive size and concomitant variables.

Since the testing of such a lattice of hypothesis falls completely within the context of standard multivariate normal linear theory there is no need to pursue this development of testing here.

6. Acknowledgements

The author is grateful to the Glasgow Blood Pressure Clinic for making available the radiological heart measurements of §4.1 and to Dr M. Damkjaer Nielsen of Glostrup Hospital, Copenhagen for the use of the steroid metabolite data of §4.2.

References

Aitchison, J. (1980). Distributions on the simplex. Paper to be presented at the International Summer School on Statistical Distributions in Scientific Work, Trieste, July 1980.

Aitchison, J. and Shen, S.M. (1980). Logistic-normal distributions: some properties and uses. *Biometrika 67*, to appear.

Aitchison, J. and Brown, J.A.C. (1957). *The Lognormal Distribution*. Cambridge University Press.

Darroch, J.N. and James, I.R. (1974). F-independence and null correlations of continuous, bounded-sum, positive variables. *J.R. Statist. Soc. B 36*, 467-483.

Mosimann, J.E. (1970). Size allometry; size and shape variables with characterizations of the lognormal and gamma distributions. *J. Amer. Statist. Ass. 65*, 930-945.

Mosimann, J.E. (1975a). Statistical problems of size and shape. I. Biological applications and basic theorems. In: *Statistical Distributions in Scientific Work, Vol. 2*. Patil, Kotz and Ord (Eds.), Reidel Publishing Company, Dordrecht, Holland, 187-217.

Mosimann, J.E. (1975b). Statistical problems of size and shape. II. Characterizations of the lognormal, gamma and Dirichlet distributions. In: *Statistical Distributions in Scientific Work, Vol. 2*. Patil, Kotz and Ord (Eds.), Reidel Publishing Company, Dordrecht, Holland, 219-239.

Nielsen, M.D. (1971). The measurement of urinary corticosteroid metabolites. *Workshop on the Diagnosis and Treatment of Cushing's Syndrome*. Glentofte Hospital, Copenhagen.

Rao, C.R. (1965). *Linear Statistical Inference and its Applications.* Wiley, New York.

Sprent, P. (1972). The mathematics of size and shape. *Biometrics 28,* 23-37.

14 CONCLUSION

This commentary has not only ranged over a wide spectrum of general theoretical problems such as estimation, hypothesis testing and multiple hypotheses problems, optimum experimental design, tolerance regions, decision theory and prediction but has also shown the relevance to practical problems such as diagnosis, calibration, treatment allocation, sampling inspection, economic demand analysis, analysis of compositional data. Moreover particular classes of distributions have been studied in relation to predictive analysis and their use in modelling especially in the more complex aspects of the practical problems of calibration and diagnosis. And some classes, for example the lognormal and the new logistic-normal, have been picked out for intensive treatment because of their relevance to a large number of consultative problems.

There are, of course, other methods of statistical analysis - non-parametric and distribution free methods, the search for robust methods, kernel methods of density function fitting - which are highly successful for many situations. In some of the more complex areas of modelling, however, such as decisive prediction, statistical diagnosis with problems of calibration, imprecision and uncertain diagnostic assessments, and the distributional problems of compositions, it is difficult to see how parametric statistical modelling can be replaced in the foreseeable future. This is not to imply, however,, that parametric statistical modelling has as yet all the answers. For example, in the modelling of the pattern of variability of

of multivariate binary data the available models, such as loglinear models, are still not entirely satisfactory.

The challenge to continue research in parametric statistical modelling remains.

REFERENCES IN THE COMMENTARY

AITCHISON, J. (1974). Hippocratic inference. *Bull. Inst. Math. Applic.* 10, 48-53.

_____ (1980a). Distributions on the simplex. Paper to be presented at the International Summer School on *Statistical Distributions in Scientific Work*, Trieste, July 1980.

_____ (1980b). Some distribution theory related to the analysis of subjective performance in inferential tasks. Paper to be presented at the International Summer School on *Statistical Distributions in Scientific Work*, Trieste, July 1980.

AITCHISON, J. and BROWN, J.A.C. (1954a). A synthesis of Engel curve theory. *Rev. Econ. Stud.* 22, 35-46.

_____ (1954b). On criteria for descriptions of income distribution. *Metroeconomica* 6, 88-107.

_____ (1954c). An estimation problem in quantitative assay. *Biometrika* 41, 338-43.

AITCHISON, J. and KAY, J.W. (1973). A diagnostic competition. *Bull. Inst. Math. Applic.* 9, 382-3.

AITCHISON, J. and MOORE, M.F. (1976). The analysis of decision-making performance. *Brit. J. Math. Statist. Psychol.* 29, 53-65.

AITCHISON, J., MOORE, M.F., WEST, S.A. and TAYLOR, T.R. (1973). Consistency of treatment allocation in thyrotoxicosis. *Quart. J. Med.* 167, 575-83.

AITCHISON, J. and SILVEY, S.D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika* 44, 131-43.

ANDREWS, D.F., GNANADESIKAN, R. and WARNER, J.L. (1973). Methods for assessing multivariate normality. In *Multivariate Analysis*, Vol. 3 (ed P.R. Krishnaiah). New York: Academic Press, pp.95-116.

ASHTON, W.D. (1972). *The Logit Transformation.* London: Griffin.

ATKINSON, A.C. (1969). A test for discriminating between models. *Biometrika* 56, 337-347.

_____ (1970). A method for discriminating between models. *J. R. Statist. Soc.* B32, 323-53.

BAIN, A.D. (1964). *The Growth of Television Ownership in the United Kingdom: A Lognormal Model.* Cambridge University Press.

BARTLETT, N.R. and DARROCH, J.N. (1978). Regression and correlation of bounded-sum variables. *Vistelius Commemoration Volume.*

BEGG, C.B. (1976). Statistical diagnosis. *Ph.D. dissertation,* University of Glasgow.

BLOCH, D.A. and WATSON, G.S. (1967). A Bayesian study of the multinomial distribution. *Ann. Math. Statist.* 38, 1423-35.

BONEVA, L.I., KENDALL, D.G. and STEFANOV, I. (1971). Spline transformations : three new diagnostic aids for the data analyst. *J. R. Statist. Soc.* B33, 1-72.

BOX, G.E.P. and COX, D.R. (1964). An analysis of transformations. *J. R. Statist. Soc.* B26, 211-52.

BROWN, G. and SANDERS, J.W. (1980). Lognormal genesis. To appear in *J. Appl. Prob.*

CARROLL, J.B. (1968). Word-frequency studies and the lognormal distribution. In *Proceedings of the Conference on Language and Language Behavior* (ed. E.M. Zale). New York: Appleton-Century-Crofts, pp.213-35.

CHAYES, F. (1960). On correlation between variables of constant sum. *J. Geophys. Res.* 65, 4185-93.

_____ (1962). Numerical correlation and petrographic variation. *J. Geol.* 70, 440-52.

CHAYES, F. and KRUSKAL, W. (1966). An approximate statistical test for correlations between proportions. *J. Geol.* 74, 692-702.

COX, D.R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symp.* 1, 105-23.

_____ (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B24, 406-24.

COX, D.R. and SMALL, N.J.H. (1978). Testing multivariate normality. *Biometrika* 65, 263-72.

CRAMER, J.S. (1962). *The Ownership of Major Consumer Durables.* Cambridge University Press.

DARROCH, J.N. (1969). Null correlations for proportions. *J. Math. Geol.* 1, 221-7.

DARROCH, J.N. and JAMES, I.R. (1974). F-independence and null correlations of continuous, bounded-sum, positive variables. *J. R. Statist. Soc.* B36, 467-83.

DARROCH, J.N. and RATCLIFF D. (1970). Null correlations for proportions II. *J. Math. Geol.* 2, 307-12.

_____ (1978). No-association of proportions. *J. Math. Geol.* 10, 361-8.

DUNSMORE, I.R. (1966). A Bayesian approach to classification. *J. R. Statist. Soc.* B28, 568-77.

_____ (1968). A Bayesian approach to calibration. *J. R. Statist. Soc.* B30, 396-405.

_____ (1969). Regulation and optimization. *J. R. Statist. Soc.* B31, 160-70.

ELDERTON, W.P. and JOHNSON, N.L. (1969). *Systems of Frequency Curves.* Cambridge University Press.

FERRIS, J.B., BROWN, J.J., FRASER, R., KAY, A.W., NEVILLE, A.M., O'MUIRCHEARTAIGH, I.G., ROBERTSON, J.I.S., SYMINGTON, T. and LEVER, A.F. (1970). Hypertension with aldosterone excess and low plasma-renin: preoperative distinction between patients with and without adrenocortical tumour. *The Lancet* 2, 995-1000.

GABRIEL, K.R. (1964). Simultaneous test procedures. *Ann. Math. Statist.* 35, 1400.

GEISSER, S. (1964). Posterior odds for multivariate normal classifications. *J. R. Statist. Soc.* B26, 69-76.

GOODMAN, L.A. (1964). A note on simultaneous confidence limits for cross-product ratios. *J. R. Statist. Soc.* B26, 86-102.

GUTTMAN, I. and TIAO, G.C. (1964). A Bayesian approach to some best population problems. *Ann. Math. Statist.* 35, 825-35.

HAIGHT, F.A. (1967). *Handbook of the Poisson Distribution.* New York: Wiley.

HART, P.E. (1960). Business concentration in the United Kingdom. *J. R. Statist. Soc.* A123, 50-8.

HEALY, M.J.R. (1968). Multivariate normal plotting. *Applied Statistics* 17, 157-61.

HEYDE, C.C. (1963). On a property of the lognormal distribution. *J. R. Statist. Soc.* B25, 392-3.

HILL, B.M. (1963). The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic. *J. Amer. Statist. Ass.* 58, 72-84.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press.

JOHNSON, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* 36, 149-76.

JOHNSON, N.L. and KOTZ, S. (1969). *Distributions in Statistics. Discrete Distributions.* Boston: Houghton Mifflin.

_____ (1970). *Distributions in Statistics. Continuous Univariate Distributions.* Boston: Houghton Mifflin.

_____ (1972). *Distributions in Statistics. Continuous Multivariate Distributions.* Boston: Houghton Mifflin.

KAY, J.W. (1976). Diagnosis and performance. *Ph.D. dissertation*, University of Glasgow.

KULLBACK, S. and LIEBLER, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.* 22, 525-40.

LANCASTER, H.O. (1969). *The Chi-squared Distribution.* New York: Wiley.

LAUDER, I.J. (1978). Computational problems in predictive diagnosis. *Compstat 1978*, 186-92.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* 60, 297-308.

LINDLEY, D.V. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.* 35, 1622-43.

_____ (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint.* Cambridge University Press.

LINDLEY, D.V., TVERSKY, A. and BROWN, R.V. (1974). On the reconciliation of probability assessments. *J. R. Statist. Soc.* A142, 146-80.

McALISTER, D. (1879). The law of the geometric mean. *Proc. Roy. Soc.* 29, 367.

MACGILLAVRAY, H.J. (1965). Variability of larger Foraminifera. *Proc. Koninkl. Nederl. Akademie van Wetenschappen* B68, 335-55.

MARDIA, K.V. (1970). *Families of Bivariate Distributions.* London: Griffin.

MATHIEU, J-R. (1978). Contribution a l'étude de la separabilité des hypothèses, au sens du test du χ^2 dans la théorie asymptotique. *Doctoral Thesis*, L'Université Paul-Sabatier de Toulouse

MIESCH, A.T. (1969). The constant sum problem in geochemistry. In *Computer Applications in the Earth Sciences* (ed D.F. Merriam). New York: Plenum Press, pp.161-77.

MORAN, M.A. and MURPHY, B.J. (1979). A closer look at two alternative methods of statistical discrimination. *Applied Statistics* 28, 223-32.

MOSIMANN, J.E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika* 49, 65-82.

_____ (1963). On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. *Biometrika* 50, 47-54.

MOSIMANN, J.E. (1970). Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions. *J. Amer. Statist. Ass.* 65, 930-45.

_____ (1975a). Statistical problems of size and shape. I. Biological applications and basic theorems. In *Statistical Distributions in Scientific Work* (eds G.P. Patil, S. Kotz and J.K. Ord). Dordrecht, Holland: D. Reidel Publishing Company, pp.187-217.

_____ (1975b). Statistical problems of size and shape. II. Characterizations of the lognormal, gamma and Dirichlet distributions. In *Statistical Distributions in Scientific Work* (eds G.P. Patil, S. Kotz and J.K. Ord). Dordrecht, Holland: D. Reidel Publishing Company, pp.219-39.

MURRAY, G.D. (1977). A note on the estimation of probability density functions. *Biometrika* 64, 150-1.

ORD, J.K. (1972). *Families of Frequency Distributions.* London: Griffin.

PATIL, G.P., KOTZ S. and ORD. J.K. (eds) (1975). *Scientific Distributions in Scientific Work.* Dordrecht, Holland: D. Reidel Publishing Company.

RAPOPORT, A. (1964). Comparison of four models for word-frequency distributions from normal and aphasic speakers. *Ph.D. dissertation,* University of North Carolina.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *J. Amer. Statist. Soc.* 47, 381-400.

SCHOENER, T.W. and JANZEN, D.H. (1968). Notes on environmental determinants of tropical versus temperate insect size pattern. *Amer. Naturalist* 102, 207-24.

SCHÜCKHER, F. (1966). Grain size. *Acta Polytechnica. Scand.* 54, 1-102.

SHAPIRO, S.S. and WILK, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591-611.

SIANO, D.B. (1972). The lognormal distribution function. *J. Chem. Educ.* 49, 755-7.

SILVEY, S.D. (1959). The Lagrangian multiplier test. *Ann. Math. Statist.*, 389-407.

SOMERS, H.H. (1959). *Analyse Mathématique de Langage: Lois Générales et Mesures Statistiques.* Louvain, Belgium: Editions Nauwelaerts.

STONE, R., AITCHISON, J. and BROWN, J.A.C. (1955). Some estimation problems in demand analysis. *The Incorporated Statistician* 5, 1-13.

SWE, C. (1964). The Bayesian analysis of contingency tables. *Ph.D. dissertation,* University of Liverpool.

TAYLOR, T.R., AITCHISON, J. and McGIRR, E.M. (1971). Doctors as decision makers: a computer-assisted study of diagnosis as a cognitive skill. *Brit. Med. J.* 3, 35-40.

THOM, H.C.S. (1963). Tornado probabilities. *Monthly Weather Rev.*, 730-6.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54, 426-82.

WILKS, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* 9, 60-2.